The Open University

MST121
Using Mathematics

Computer Books A–D

About MST121

This module, MST121 *Using Mathematics*, and the modules MU123 *Discovering mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. See the address below for further details.

MST121 uses Mathcad (Parametric Technology Corporation) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of MST121.

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 60 90, email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom, for a brochure (tel. +44 (0)1908 858793, fax +44 (0)1908 858787, email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2010.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed and bound in the United Kingdom by Cambrian Printers, Aberystwyth.

The paper used in this publication is procured from forests independently certified to the level of Forest Stewardship Council (FSC) principles and criteria. Chain of custody certification allows the tracing of this paper back to specific forest-management units (see www.fsc.org).

ISBN 978 1 8487 3352 7

1.1

Contents

Computer Book A

Mathematics and Modelling

Guidance notes

This computer book contains those sections of the chapters in Block A which require you to use Mathcad. Each of these chapters contains instructions as to when you should first refer to particular material in this computer book, so you are advised not to work on the activities here until you have reached the appropriate points in the chapters.

In order to use this computer book, you will need the following Mathcad files.

Chapter A1

121A1-01	Sequences in Mathcad
121A1-02	Arithmetic sequences
121A1-03	Geometric sequences
121A1-04	Linear recurrence sequences
121A1-05	The mortgage sequence (Optional)

Chapter A2

121A2-01	Parametric equations of lines
121A2-02	Parametric equations of circles
121A2-03	Compound circular motion (Optional)

Chapter A3

121A3-01	Functions graphs and solutions
121A3-02	The Mathcad solve block
121A3-03	Solving equations symbolically
121A3-04	Mathcad graph plotter

Instructions for installing these files onto your computer's hard disk, and for opening them, are given in Chapter A0.

Activities based on software vary both in nature and in length. Sometimes the instructions for an activity appear only in the computer book; in other cases, instructions are given in the computer book and on screen. Feedback on an activity is sometimes provided on screen and sometimes given in the computer book.

For advice on how each computer session fits into suggested study patterns, refer to the Study guides in the chapters.

Chapter A1, Section 6
Investigating sequences with the computer

In this section, you will use the computer to investigate the mathematical behaviour of arithmetic sequences, geometric sequences and linear recurrence sequences. The computer enables many terms to be computed quickly, graphs to be plotted easily and the long-term behaviour of sequences to be observed directly: by just changing a parameter, the terms of the sequence are recalculated and the graph redrawn. The long-term behaviour of linear recurrence sequences was discussed in Section 5, where the notation $a_n \to \infty$ as $n \to \infty$, and $a_n \to l$ as $n \to \infty$, was introduced.

There are five Mathcad files accompanying this section. The first file shows how to set up, display and plot graphs of sequences in Mathcad. The next three files cover the important mathematics of this section; they all have a similar structure and contain investigations of the behaviour of arithmetic, geometric and linear recurrence sequences, respectively. The final file, on the mortgage sequence, is entirely optional.

Assistance and explanations are provided on the pages within each file, with detailed Mathcad instructions where appropriate. In addition, help is available from the following.

◇ The MST121 reference manual, *A Guide to Mathcad*.
This contains full descriptions of all the Mathcad commands used in the files, including information about Mathcad error messages.

◇ Mathcad's own on-screen help facility.
Select the **Help** menu and **Mathcad Help**, then choose the 'Contents' tab for details of basic Mathcad operations, or the 'Index' or 'Search' tabs to search for a particular topic. (Please note that this information is not specific to MST121.)

6.1 Arithmetic sequences

The book sequence

See Chapter A1, Section 2.

Recall that the book sequence,

$$5, 8, 11, 14, \ldots, 38,$$

represents the total number of books that have been received by the end of successive months by a member of a book club. The closed form of this arithmetic sequence is

$$b_n = 5 + 3(n - 1) \quad (n = 1, 2, 3, \ldots, 12).$$

In Activity 6.1, you are invited to set up this sequence on the computer, and plot a graph of it. This Mathcad file may look quite long, but all the tasks are straightforward.

Activity 6.1 Setting up sequences in Mathcad

Remember to make your own working copy of the file; see Chapter A0, Subsection 2.2, for guidance.

Locate in the **Chapter A1** folder the file **121A1-01 Sequences in Mathcad**, and open it. Page 1 introduces the worksheet. Read pages 2 to 8, follow the instructions, and carry out Tasks 1 to 5.

Comment

- ◇ Tasks 1 and 2 introduce subscripted variables and range variables, which play essential roles in setting up sequences in Mathcad.
- ◇ Task 3 involves defining a range variable and 12 subscripted variables to create the book sequence, and then using a table of values to display the sequence by entering $b =$. Such tables will be used throughout MST121. Their appearance depends on the number of terms in the sequence – details of the different forms that tables can take are given in *A Guide to Mathcad*.

  Note that the Mathcad expressions can easily be modified if the rules of the book club change. For example, a change to sending four books each month, rather than three, could be handled by editing the definition for the subscripted variable $b_n := 5 + 3(n - 1)$ to read $b_n := 5 + 4(n - 1)$.
- ◇ Tasks 4 and 5 involve creating a graph and then formatting the graph display. You will meet these Mathcad techniques throughout MST121 – once again, the range variable plays an important part.

  Note that Mathcad calls the horizontal and vertical axes the x- and y-axes, although the graph is actually a plot of b_n on the vertical axis against n on the horizontal axis. In fact, Mathcad refers to the whole graph as an 'X-Y Plot'. (This type of graph traditionally plots 'y against x', but any two variables can be plotted.)

Now close Mathcad file 121A1-01.

A general arithmetic sequence – varying the parameters

A general arithmetic sequence with recurrence system

See Chapter A1, Section 2.

$$x_0 = a, \quad x_{n+1} = x_n + d \quad (n = 0, 1, 2, \ldots),$$

where a is the first term and d is the common difference, has closed form

$$x_n = a + nd \quad (n = 0, 1, 2, \ldots).$$

Here, as with all the general sequences in this section, we start with the term x_0, rather than x_1, since this results in a simpler closed form. We obtain the *same terms* for the sequence, but with the subscripts decreased by one.

In particular, whether we start with x_0 or x_1 does not affect the long-term behaviour of the sequence.

Choosing the values of the parameters a and d determines a particular arithmetic sequence. In the next activity you will use Mathcad to investigate the effects of changing these parameters. Altering the parameters *one at a time* allows you to distinguish between the effects of changing each of them.

Activity 6.2 Investigating arithmetic sequences

Open Mathcad file **121A1-02 Arithmetic sequences**. Page 1 introduces the worksheet. Look at page 2, where the task starts with the arithmetic sequence obtained by setting $a = 10$ and $d = 1$, namely

$$x_n = 10 + n \quad (n = 0, 1, 2, \ldots, 20).$$

(a) Investigate the effect on the sequence of changing the parameter a, while keeping the parameter d constant (leave $d = 1$). Use the following values for a in turn:

$$10, \quad 0, \quad -20.$$

In an investigation of this type, it helps to choose round numbers and to change them in a systematic way, for example, from positive to negative values.

In each case, fill in the corresponding cell of the $d = 1$ row of the table in Figure 6.1 with a small sketch graph, and add a comment similar to the ones already filled in.

State briefly how altering a affects the graph of the sequence.

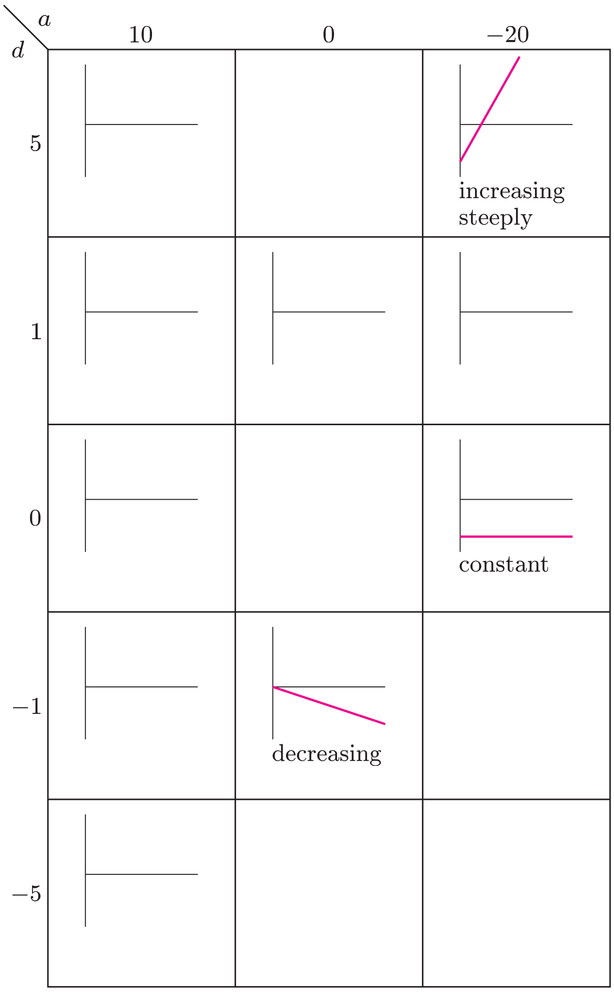


Figure 6.1 Sketch graphs of arithmetic sequences

The sketch graphs here are intended to indicate the *overall* shape of the graph. We therefore draw line graphs, rather than plotting individual terms of the sequences.

(b) Investigate the effect on the sequence of changing the parameter d, while keeping the parameter a constant. Set $a = 10$. Then use the following values for d in turn:

$$5, \quad 0, \quad -1, \quad -5.$$

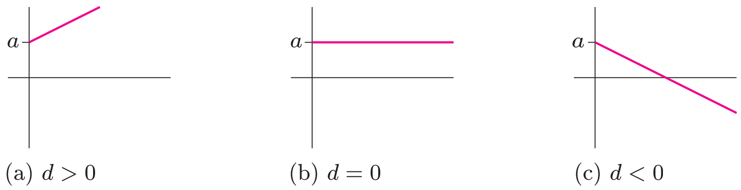
In each case, fill in the corresponding cell of the $a = 10$ column of the table in Figure 6.1.

State briefly how altering d affects the graph of the sequence.

Solutions are given on page 34.

Comment

The graph of an arithmetic sequence lies along a straight line. The graph can have one of the basic shapes in Figure 6.2. Each graph shown has $a > 0$.



(a) $d > 0$ (b) $d = 0$ (c) $d < 0$

Figure 6.2 Basic shapes of graphs of arithmetic sequences

The parameter d determines the slope of the graph (upwards or downwards).

- If $d > 0$, then the graph slopes upwards (from left to right); the terms of the sequence are increasing.
- If $d = 0$, then the graph is horizontal; the terms of the sequence are constant.
- If $d < 0$, then the graph slopes downwards; the terms of the sequence are decreasing.

These statements remain true if the first term is $x_1 = a$ and the common difference is d.

The larger d is (positive or negative), the steeper is the slope of the graph.

The parameter a determines the first term of the sequence. If $a > 0$, then the first term is above the horizontal axis (as shown in Figure 6.2 above); if $a = 0$, then the first term lies on the horizontal axis; and if $a < 0$, then the first term lies below the horizontal axis.

Mathcad notes

Mathcad notes provide extra information about the features and techniques used in the Mathcad files. They are *optional*.

- The built-in Mathcad variable *ORIGIN* specifies the subscript used for the first term of a sequence. By default, Mathcad sets *ORIGIN* to zero for every worksheet. Since the sequence x_n here has first term x_0, there is no need to alter the value of *ORIGIN*.
- When displaying sequences which have many terms, Mathcad provides a scrolling table of values. To access all of the values, click once on a value in the table to reveal the scroll bar.
- The graph has been resized to make it appear larger on the page.
- The vertical (y-) axis scale has been fixed from -50 to 50.
- The graph trace has been formatted to display the individual terms of the sequence as blue crosses. On the vertical (y-) axis 'Auto grid' has been switched off, and 'Number of grids' has been set to 10, to improve the axis labelling. In addition, the 'Show markers' option has been switched on for the vertical (y-) axis, with a dashed red marker line drawn at $y = 0$ to indicate the horizontal (x-) axis.

Resizing a graph and fixing the graph scale are discussed in Chapter A2.

These features are set by formatting the graph: first choose **Graph** and **X-Y Plot** from the **Format** menu, then select the 'X-Y Axes' tab from the resulting option box.

Now close Mathcad file 121A1-02.

6.2 Geometric sequences

Remember that we are starting with the term x_0, because this gives a simpler closed form.

A general geometric sequence with recurrence system

$$x_0 = a, \quad x_{n+1} = rx_n \quad (n = 0, 1, 2, \ldots),$$

where a is the first term and r is the common ratio, has closed form

$$x_n = ar^n \quad (n = 0, 1, 2 \ldots).$$

Notice that for $r = 0$ this closed form gives the correct value $x_0 = a$ only if we adopt the convention that $0^0 = 1$.

Choosing the values of the parameters a and r determines a particular geometric sequence. The long-term behaviour of the sequence x_n depends on the long-term behaviour of the sequence r^n, which is summarised in Table 6.1 for various ranges of r, including negative ones.

See Chapter A1, Section 5.

Table 6.1 Long-term behaviour of r^n

Range of r	Behaviour of r^n
$r > 1$	$r^n \to \infty$ as $n \to \infty$
$r = 1$	Remains constant: 1, 1, 1, ...
$0 < r < 1$	$r^n \to 0$ as $n \to \infty$
$r = 0$	Remains constant: 0, 0, 0, ...
$-1 < r < 0$	$r^n \to 0$ as $n \to \infty$, alternates in sign
$r = -1$	Alternates between -1 and $+1$
$r < -1$	Unbounded, alternates in sign

In the next activity, you will use Mathcad to investigate the effects of changing the parameters a and r of a geometric sequence. In particular, you will observe the long-term behaviour of r^n given in Table 6.1.

Activity 6.3 Investigating geometric sequences

Open Mathcad file **121A1-03 Geometric sequences**. Look at page 2 of the worksheet, where the task starts with the geometric sequence obtained by setting $a = 1$ and $r = 2$, namely

$$x_n = 2^n \quad (n = 0, 1, 2, \ldots, 20).$$

The graph on page 2 of the worksheet has a fixed vertical scale; only the terms of the sequence with values lying between -5 and 5 are shown. You can see from the table of values that the terms of this particular sequence quickly exceed 5. (Note that this table contains all of the values of the sequence – click on a value in the table to reveal the scroll bar.)

To move quickly between pages of the worksheet, press `[Shift][Page Up]` and `[Shift][Page Down]`.

Now look at the graph on page 3 of the worksheet. This is a graph of the same sequence, but it rescales automatically so that the largest term (positive or negative) is always shown. This graph shows the long-term behaviour of the sequence, even if the later terms are very large (as is the case when $a = 1$ and $r = 2$), or very small. When interpreting sequence behaviour from this graph, it is important to look at the scale on the vertical axis.

Large numbers on the scale are displayed using scientific notation; for example, 500 000 is shown as 5×10^5.

You may find it helpful to use both graphs in the activity, but it is best to base your sketch graphs on the one on page 2, which allows direct comparisons to be made between graphs of different sequences. In fact, you need to be on page 2 in order to alter the parameters a and r, so you should return there now.

The sequence r^n grows very quickly when $r \geq 2$. We therefore choose values of r close to 1 in this study of the behaviour of geometric sequences.

(a) Investigate the effect on the sequence of changing the parameter a, while keeping the parameter r constant. Set $r = 1.1$. Then use the following values for a in turn:

$$2, \quad 1, \quad -1.$$

In each case, fill in the corresponding cell of the $r = 1.1$ row of the table in Figure 6.3.

State briefly how altering a affects the graph of the sequence.

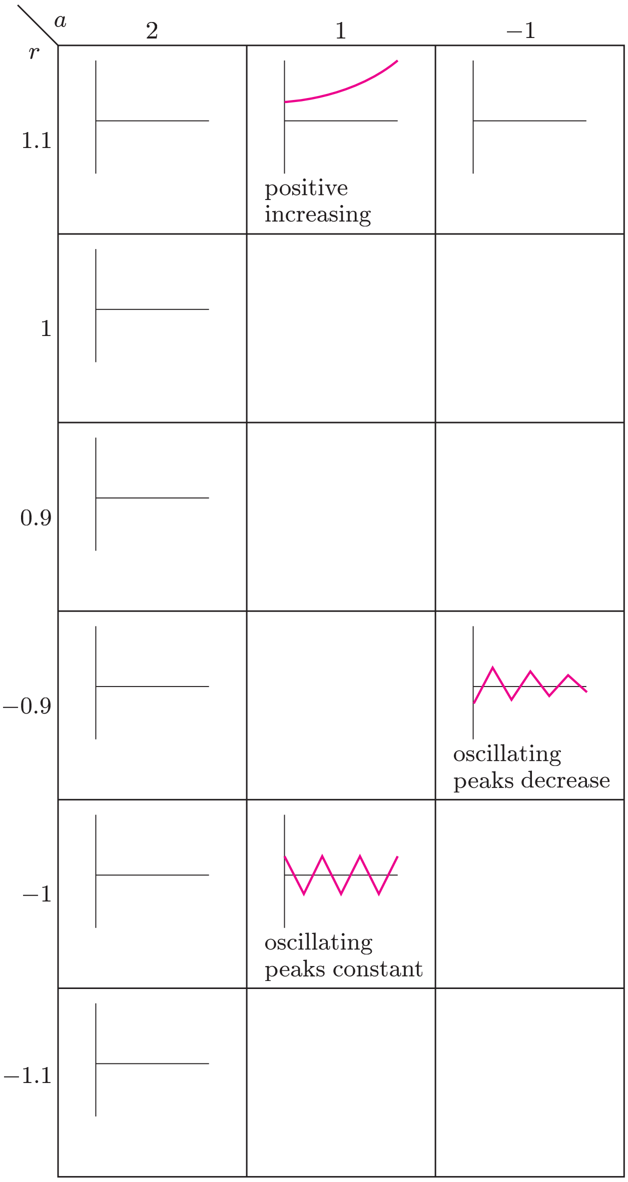


Figure 6.3 Sketch graphs of geometric sequences

Remember that these sketch graphs are intended to indicate the *overall* shape of the graph.

(b) Investigate the effect on the sequence of changing the parameter r, while keeping the parameter a constant. Set $a = 2$. Then use the following values for r in turn:

$$1, \quad 0.9, \quad -0.9, \quad -1, \quad -1.1.$$

In each case, fill in the corresponding cell of the $a = 2$ column of the table in Figure 6.3. Try to predict the results before changing the parameter in the Mathcad worksheet.

State briefly how altering r affects the graph of the sequence.

Solutions are given on page 34.

Comment

The graph of a geometric sequence can have one of the basic shapes in Figure 6.4. Each graph shown has $a > 0$. The shapes for $a < 0$ are obtained by reflecting these graphs in the horizontal axis; that is, flipping them upside down.

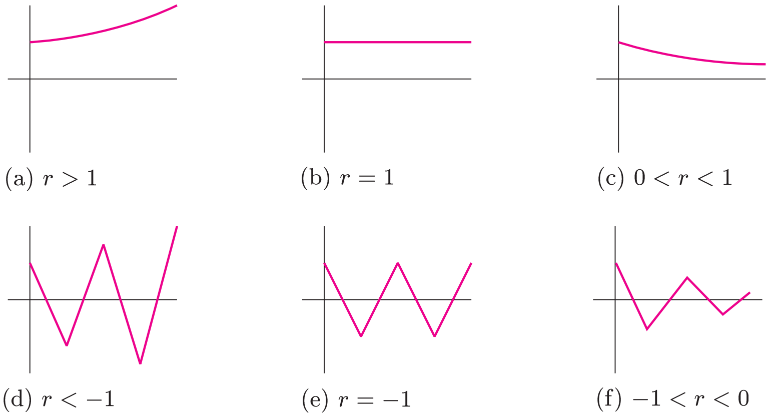


Figure 6.4 Basic shapes of graphs of geometric sequences

The parameter r determines the overall shape of the graph – whether it is constant, tends to infinity, flattens out, or oscillates with constant, increasing or decreasing peaks. The parameter a determines the first term of the sequence. If $a = 0$, then the graph is constant with all the terms equal to zero.

Mathcad notes

In mathematics, the expression 0^0 is not generally defined, except where a convention is used. Mathcad, however, gives $x^0 = 1$ for *all* real x.

◇ If you change r to 0, then Mathcad gives the correct value for the first term, $x_0 = a$. When $r = 0$ and $n = 0$, Mathcad evaluates r^n to give 1.

◇ The result format has been set to ‘Number of decimal places’ 6 and ‘Exponential threshold’ 9. These settings are used for all the ‘results’ in the worksheet. However, Mathcad does not use these settings to display the numbers on the graph scales. (To observe the number settings for graphs, choose **Graph** and **X-Y Plot** from the **Format** menu, then select the ‘Number Format’ tab from the resulting option box.)

◇ To make it easier to follow the order of the sequence, both graphs have been formatted to join the individual points (blue crosses) with a dotted line. To do this, the graph trace has been set with ‘Symbol’ ×, dotted ‘Line’, blue ‘Color’ and ‘Type’ lines.

◇ The graph on page 3 always rescales automatically. Such automatic rescaling, in order to plot all the data, is Mathcad’s default.

Now close Mathcad file 121A1-03.

6.3 Linear recurrence sequences

A general linear recurrence sequence is given by

See Chapter A1, Section 4.

$$x_0 = a, \quad x_{n+1} = rx_n + d \quad (n = 0, 1, 2, \ldots).$$

The overall shape of the graph of the sequence depends on the value of r. If $r \neq 1$, then the sequence has the closed form

$$x_n = \left(a + \frac{d}{r-1}\right) r^n - \frac{d}{r-1} \quad (n = 0, 1, 2, \ldots), \tag{6.1}$$

which is the sum of the geometric sequence

$$\left(a + \frac{d}{r-1}\right) r^n$$

and the constant sequence

$$-\frac{d}{r-1} = \frac{d}{1-r}.$$

The graph of such a linear recurrence sequence must therefore have one of the basic shapes shown in Figure 6.4, but displaced vertically by the amount $d/(1-r)$. On the other hand, a linear recurrence sequence with $r = 1$ is an arithmetic sequence with one of the basic shapes shown in Figure 6.2.

A particular linear recurrence sequence is obtained by assigning values to the parameters a, r and d. A systematic investigation of the effects of changing all *three* parameters would result in a great number of cases to consider, not to mention the problem of displaying the results in a three-dimensional table! We consider just one type of behaviour, where $0 < r < 1$ so $r^n \to 0$, and investigate the effect on this type of sequence of altering the other parameters, a and d.

Activity 6.4 Investigating linear recurrence sequences

Open Mathcad file **121A1-04 Linear recurrence sequences**. The layout of this worksheet is identical to that of the previous one, with the table and fixed-scale graph on page 2, and the graph that automatically rescales on page 3. There is one minor change in this worksheet, however. The parameter N allows you to change the number of terms of the sequence that are displayed. You can increase N if you feel that this would be helpful when determining the long-term behaviour, though it should be possible to see all the effects of changing a, r and d with the value of N provided.

Look at page 2, where the task starts with the linear recurrence sequence obtained by setting $a = 1$, $r = 0.5$ and $d = 1$, namely

$$x_n = -(0.5)^n + 2 \quad (n = 0, 1, 2, \ldots, 20).$$

(a) Investigate the effect on the sequence of changing the parameter a, while keeping the parameters r and d constant (leave $r = 0.5$ and $d = 1$). Use the following values for a in turn:

To change the parameters, you need to be on page 2.

$4, \quad -4.$

In each case, fill in the corresponding cell of the $d = 1$ row of the table in Figure 6.5.

State briefly how altering a affects the graph of the sequence.

Remember that these sketch graphs are intended to indicate the *overall* shape of the graph.

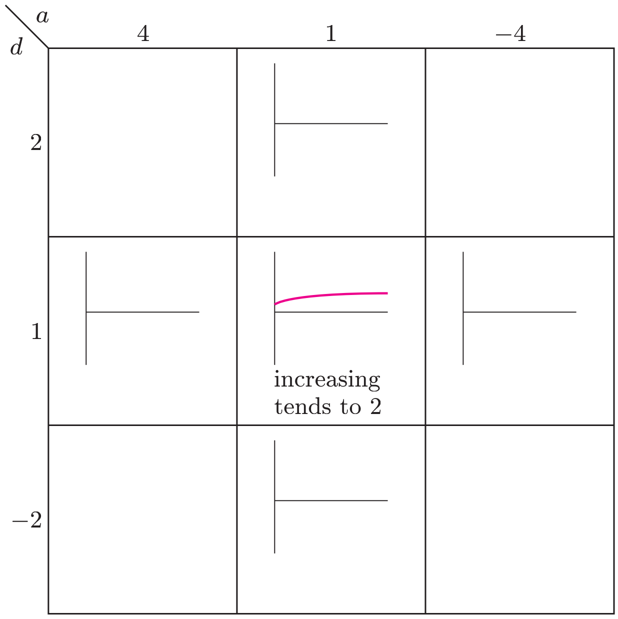


Figure 6.5 Sketch graphs of linear recurrence sequences with $r = 0.5$

(b) Investigate the effect on the sequence of changing the parameter d, while keeping the parameters a and r constant (set $a = 1$ and leave $r = 0.5$). Use the following values for d in turn:

$2, \quad -2.$

In each case, fill in the corresponding cell of the $a = 1$ column of the table in Figure 6.5.

State briefly how altering d affects the graph of the sequence.

Solutions are given on page 35.

Comment

The parameter a has no effect on the long-term behaviour, but it does affect the initial term. The reverse is true for the parameter d. For $-1 < r < 1$, the graph of the sequence flattens out or the peaks get smaller and smaller, and the sequence tends to $d/(1 - r)$, which is the constant term in the closed form of equation (6.1).

Mathcad notes

In mathematics, the expression $0/0$ is not defined. Mathcad, however, gives $0/x = 0$ for *all* real x.

◇ If you change the parameter r to 1, so $r - 1 = 0$, then Mathcad will give no terms for the sequence and will not plot a graph, except when the parameter d is 0. When $d = 0$, Mathcad evaluates the subexpression $d/(r - 1)$ in the closed form to give 0. When $d \neq 0$, Mathcad is unable to evaluate this, and gives an error.

◇ Using a variable N to specify the number of terms in the sequence aids clarity and makes it easier to change things. Once N has been defined, it can be used as the final value in the definition for the range variable $n := 0, 1 \,..\, N$ (on page 2) and to display the last term of the sequence as x_N (on page 3). Both of these can then be altered by simply changing the value in the definition for N. Note that, in Mathcad, N and n are two different variables. Note also that the range variable n runs from 0 to N, so we are actually calculating $N + 1$ terms of the sequence x_n.

Now close Mathcad file 121A1-04.

We conclude this section with the following optional activity.

Activity 6.5 Mortgage repayments (Optional)

You may like to use Mathcad file **121A1-05 The mortgage sequence** to experiment with repayments on a loan. The worksheet starts with the mortgage sequence, representing a loan of £10 000 to be repaid over a period of 20 years at an annual interest rate of 5%. You may like to try changing the size of the initial loan £L, the interest rate R%, the annual repayment £P, or the term of the loan N years.

See Chapter A1, Section 4.

Now close Mathcad file 121A1-05.

Chapter A2, Section 5
Parametric equations by computer

In this section, you will use the computer to plot lines and circles described by parametric equations.

There are three Mathcad files accompanying this section. The first and second files plot lines and circles, respectively. The final file is optional; it involves plotting more complicated curves that arise from compound circular motion.

5.1 Parametric equations of lines

See Chapter A2, Subsection 4.1.

Recall that a line $y = mx + c$, with slope m and y-intercept c, can be described by a pair of parametric equations

$$x = t, \quad y = mt + c.$$

This parametrisation can be used to plot the line in Mathcad, for particular values of m and c. The key requirement is to define a range variable to represent the parameter t. The graph is then constructed by entering the expressions describing x and y in terms of this range variable on the horizontal (x-) axis and vertical (y-) axis, respectively. This approach provides a powerful graphing technique in Mathcad, which can be used to plot *any* curve described by a pair of parametric equations.

Activity 5.1 Plotting a line described by parametric equations

Open Mathcad file **121A2-01 Parametric equations of lines**. Read page 2 of the worksheet, and carry out Task 1.

Solutions and some comments are included in the Mathcad worksheet.

See Chapter A2, Section 1.

Given two lines, it is possible to determine where they intersect using algebraic techniques. In the next activity, you are asked to find *approximately* where two lines intersect by identifying the intersection point on the graph using Mathcad's graph trace tool.

Activity 5.2 Finding intersections of two lines graphically

You should still be working with Mathcad file 121A2-01.

Read pages 3 and 4 of the worksheet, and carry out Tasks 2 and 3.

The solutions are provided in the worksheet as part of the comments following each task.

Comment

It is difficult to determine the *exact* point of intersection of two lines using this method. However, it is possible to find small ranges within which the coordinates of the point of intersection lie. For example, the lines in Task 3, with parametric equations $x = t$, $y = 2t + 3$ and $x = 4t - 7$, $y = -t + 16$, do intersect. This point of intersection has x-coordinate between 4.5 and 5.5, and y-coordinate between 12.5 and 13.5.

Note that the graph range has been chosen to include the point of intersection.

Mathcad notes

The techniques used to plot two curves on a graph (the two lines on page 3 of the worksheet) can be extended to plot three, four or more curves on the same graph. Simply enter all the expressions for 'x' separated by commas on the horizontal (x-) axis, and likewise the 'y' expressions on the vertical (y-) axis. Remember that Mathcad matches the expressions in pairs – the first 'y' expression is plotted against the first 'x' expression, the second against the second, and so on.

Remember that Mathcad notes are *optional*.

Note that a line is a type of curve, even though it is not 'curved'.

In the main text, you saw that parametric equations for lines provide a straightforward method for determining the closest approach of two objects in linear motion. In the next *optional* activity, the computer is used to illustrate the tracing out of paths with respect to time, from which it can be seen whether two ships following intersecting courses actually collide.

See Chapter A2, Example 4.2 and Activity 4.3.

Activity 5.3 Collision course? (Optional)

Look at page 5 of the worksheet, and follow the instructions.

You should still be working with Mathcad file 121A2-01.

Notice that the graph plots the two lines (the courses of the ships) and the right-hand endpoints of the lines (the positions of the ships), so there are four expressions on each axis. For a given value of T, the graph plots:

- the line $x = t$, $y = 2t + 3$ in red, and the line $x = 4t - 7$, $y = -t + 16$ in blue, both for the range $0 \leq t \leq T$;
- the point $(T, 2T + 3)$ as a red box symbol, and the point $(4T - 7, -T + 16)$ as a blue diamond symbol.

Mathcad notes

The square root symbol, used when calculating the distance between the two ships, can be obtained from the appropriate button on the 'Calculator' toolbar, or by typing \ (backslash). For example, you can type \2= to obtain $\sqrt{2} = 1.414$ (to 3 d.p.).

Now close Mathcad file 121A2-01.

5.2 Parametric equations of circles

Recall that a circle, with centre at (a, b) and radius r, has parametrisation

$$x = a + r\cos t, \quad y = b + r\sin t \quad (0 \leq t \leq 2\pi).$$

See Chapter A2, Subsection 4.2.

This parametrisation is ideal for plotting circles using Mathcad.

Activity 5.4 Plotting circles from parametric equations

Open Mathcad file **121A2-02 Parametric equations of circles**. Read page 2 of the worksheet, and carry out Task 1.

Solutions are given on page 3 of the worksheet.

Comment

You might have been surprised that none of the circles you obtained actually looks very circular! This is due to the way that Mathcad scales and sizes the graphs. All the curves do in fact represent circles, as you will see in Activity 5.5.

Mathcad notes

Both sine and cosine are built-in functions in Mathcad, and by default they work with angles in radians. They are obtained from the appropriate buttons on the 'Calculator' toolbar or by typing `sin` (or `cos`). The 'sin' and 'cos' buttons automatically insert a pair of round brackets after the function, that is, a left bracket, followed by an empty placeholder and a right bracket. Entry continues in this placeholder, and you can enter the value to which sin (or cos) is to be applied to obtain, for example, $\sin(t)$ or $\cos(\pi)$. If you wish to enter further information outside the brackets, then you will need to press `[Space]` to extend the editing lines around the brackets (and function) before you can do so. When entering the functions via the keyboard, you must type `sin` and `cos` using lower-case letters, and you must also enter the brackets yourself, for example, type `sin(t)` or `cos([Ctrl][Shift]p)`.

When is a circle not a circle?

As you saw in Activity 5.4, a circle plotted by Mathcad may not always look like a circle. Consider the graphs below, all plotted over the range $t := 0, 0.01 .. 2\pi$. Which of these are actually the graphs of circles?

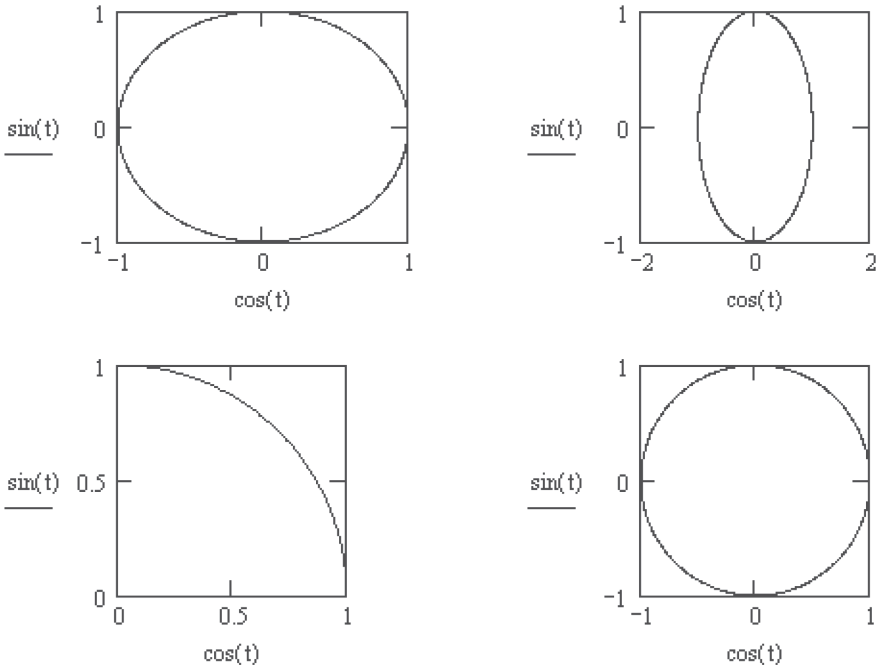


Figure 5.1 Circles plotted by Mathcad

Each of the graphs arises from these parametric equations of the unit circle:

$$x = \cos t, \quad y = \sin t \quad (0 \leq t \leq 2\pi).$$

The first graph is the default graph plotted by Mathcad – the curve is stretched horizontally to fit the rectangular graph box, and therefore appears to be an oval. The other three graphs have been resized to make their graph boxes square. They illustrate the effect of different scales on the appearance of a graph. The top right graph has the horizontal (x-) axis scale set from -2 to 2, while the vertical (y-) axis scale is set from -1 to 1. This results in the circle appearing to be squashed. The bottom left graph has both the horizontal (x-) axis and the vertical (y-) axis scales set from 0 to 1. This results in only one quarter of the circle being displayed. The final graph shows a circle which looks circular – the graph box is square and the scales of both axes are from -1 to 1, the default scales, thus showing the whole circle undistorted.

The mathematical name for a stretched circle is an *ellipse*. Ellipses are studied in MS221.

Note that Mathcad plots one point for each value of the range variable, and then joins them together with line segments. Therefore the step size chosen when defining the range variable plays a crucial role in the appearance of the curve plotted. For example, consider the following graph of the unit circle, where the range variable is set to $t := 0, 1 \,..\, 2\pi$.

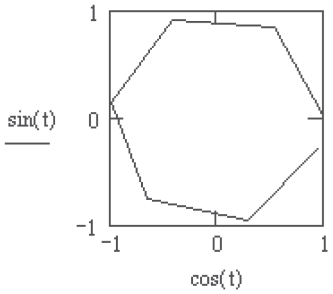


Figure 5.2 Mathcad plot of a circle where the step size is too large

As a rule of thumb, for a graph that is not a straight line, Mathcad should plot at least 100 points to be sure of obtaining a satisfactory picture; for example, the step size could be 0.1 for the range 0 to 10, and 0.05 for the range 0 to 2π.

These examples show that appearances may be deceptive. Mathcad always chooses a scale that allows it to display the whole curve, unless the scale is fixed manually. This may result in the graph being shorter and fatter, or taller and thinner, than the true picture. In the next activity you can practise resizing and fixing the scales of graphs to make them appear as you would expect.

Activity 5.5 How to resize a graph and fix the graph scale

Read pages 4 and 5 of the worksheet, and carry out Task 2.

You should still be working with Mathcad file 121A2-02.

Mathcad notes

◇ The instructions on how to fix the graph scale (on page 4 of the worksheet) relate to fixing the scale after a graph has been drawn. You can also fix the scale while you create a graph, when the graph box and all the placeholders are empty. (The four empty placeholders for the limits of the axes appear at the ends of the horizontal and vertical axes.)

◇ When formatting graphs, you may have noticed an 'Auto scale' option in the axis settings. This option applies only when Mathcad automatically sets the axis limits, and has no effect if you fix the graph scale by setting the axis limits yourself. When 'Auto scale' is on (the default), Mathcad sets the axis limits to round numbers. When it is off, Mathcad sets the axis limits to the extreme values of the data, so the traces plotted on the graph extend to the edges of the graph box.

The parametrisation representing motion in a circular path is

See Chapter A2, Subsection 4.2.

$$x = a + r\cos(kt), \quad y = b + r\sin(kt) \quad (T_1 \leq t \leq T_2),$$

where (a, b) is the centre of the circle and r is its radius. The non-zero constant k determines the rate and direction (anticlockwise or clockwise) of the motion, and the parameter range endpoints, T_1 and T_2, determine the portion of the circle traversed.

In the next activity you are asked to investigate the effects of altering the values of a, b and r on the position and size of the circle plotted, and the effects of altering the values of k, T_1 and T_2 on the portion of the circle plotted.

Activity 5.6 Exploring the parametrisation of a circle

You should still be working with Mathcad file 121A2-02.

Read page 6 of the worksheet, and carry out Task 3.

Mathcad notes

We can avoid subscripts in Mathcad by using the variables $T1$ and $T2$ rather than T_1 and T_2.

The range variable t is defined as $t := T1, T1 + 0.01 \,..\, T2$. Thus t ranges from $T1$ to $T2$ in steps of size 0.01, where the values of $T1$ and $T2$ have already been defined. When defining range variables, it is important to remember that the three numbers in the definition are

starting value, next value .. final value.

The next activity involves parametric equations of both lines and circles. You will use the techniques that you have learned so far to find approximations to the points of intersection of a given line and circle.

Activity 5.7 Finding points of intersection graphically

You should still be working with Mathcad file 121A2-02.

Read page 7 of the worksheet, and carry out Task 4.

A solution is given on page 8 of the worksheet.

Comment

Notice that we needed to define *two* range variables here. The range variable t is used for the parametrisation of the circle, and the range variable u is used for the parametrisation of the line.

The letter most commonly used for a parametrisation is t. When other parameters are required, the letters s, u and v are often used.

Now close Mathcad file 121A2-02.

5.3 Two-circle compound motion (Optional)

Two-circle compound motion describes the position of a point determined simultaneously by *two* circles. Consider, for example, an electric food mixer with various attachments; see Figure 5.3. The point of attachment rotates about the central axis of the mixer (circle 1), and the attachment itself rotates about this point of attachment (circle 2).

The motion of a point on the edge of an attachment has the following type of parametrisation:

$$x = r\cos(kt) + R\cos(Kt), \quad y = r\sin(kt) + R\sin(Kt). \tag{5.1}$$

The values of r and k relate to circle 1, with r being the radius and k determining the rate of rotation. The values of R and K relate to circle 2 in a similar manner.

In a particular food mixer, an attachment rotates $3\frac{1}{3}$ times about the point of attachment every time this point rotates once about the central axis. So the ratio of K to k is 10:3. The radius r is a fixed distance, whereas R depends on the attachment.

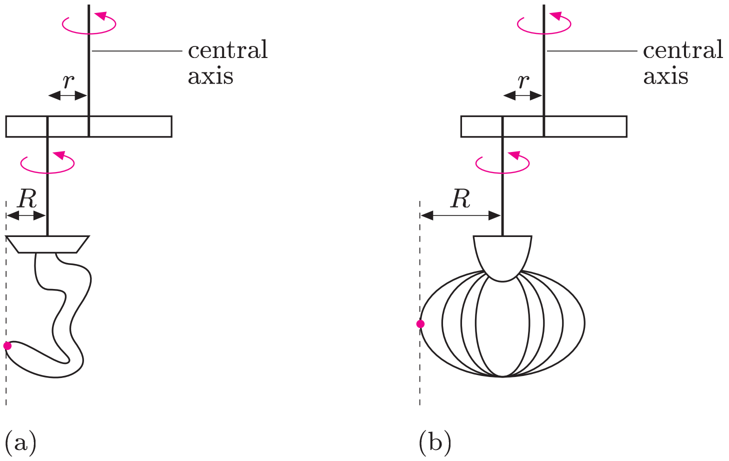
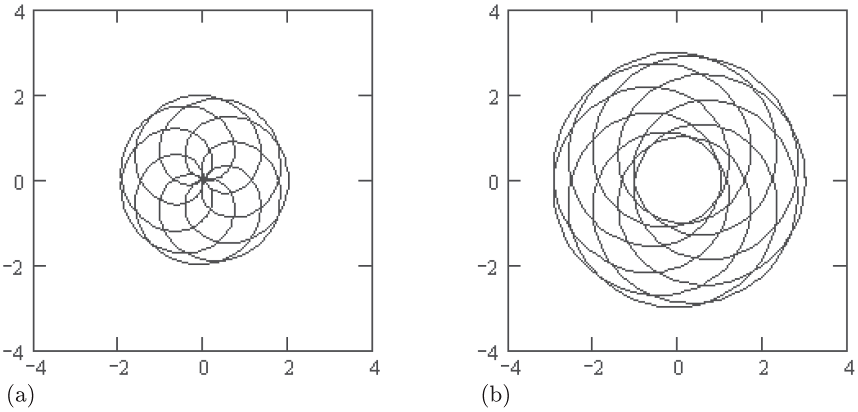


Figure 5.3 (a) Dough hook: $R = r$ (b) Whisk: $R = 2r$

On setting $r = 1$, $k = 3$, $R = 1$ and $K = 10$ in equation (5.1) for the dough hook, and $r = 1$, $k = 3$, $R = 2$ and $K = 10$ for the whisk, we see that the points plotted trace out the following curves.



Notice that the curve for the dough hook passes through the origin, but the curve for the whisk does not.

Figure 5.4 Curves traced by points on attachments: (a) dough hook, (b) whisk

You can create such graphs, associated with two-circle compound motion, in the final activity.

Activity 5.8 Compound circular motion (Optional)

Open Mathcad file **121A2-03 Compound circular motion**. Read the worksheet, and alter the values of r, k, R and K to investigate what patterns you can obtain.

Comment

If you wish to see how the dough hook and whisk patterns are built up by plotting just one revolution, then set

$$r = 1, \quad k = 1, \quad R = 1 \quad \text{and} \quad K = 10/3 \quad \text{for the dough hook,}$$
$$r = 1, \quad k = 1, \quad R = 2 \quad \text{and} \quad K = 10/3 \quad \text{for the whisk.}$$

Now close Mathcad file 121A2-03.

Chapter A3, Section 5
Functions, graphs and equations on the computer

In this section, you will use various Mathcad techniques to solve problems involving functions, graphs and equations.

There are four Mathcad files that accompany this section. In each of the first three files a different technique is introduced, which is used first to solve the exhibition hall problem and then to solve a similar but more complicated problem. The fourth file is set up as a general graph plotter. This can be used to plot graphs of the functions introduced in the main text, and also to solve two further problems concerning the *least* values taken by particular functions.

See Chapter A3, Subsection 2.1.

5.1 Problems and equations

First, here are three problems and their associated equations and functions.

The exhibition hall problem

Recall that the exhibition hall problem leads to the equation

$$4x^2 - 56x + 192 = 96. \tag{5.1}$$

Here x is the width of the border, the expression $4x^2 - 56x + 192$ represents the area of clear space in the exhibition hall, and 96 is half the total area. We represent the area of clear space using the function

All lengths used here are in metres.

$$f(x) = 4x^2 - 56x + 192 \quad (x \text{ in } [0,6]), \tag{5.2}$$

so equation (5.1) can be written in the form $f(x) = 96$. Figure 5.1 shows the graph of $y = 4x^2 - 56x + 192$. The solid part is the graph of $y = f(x)$.

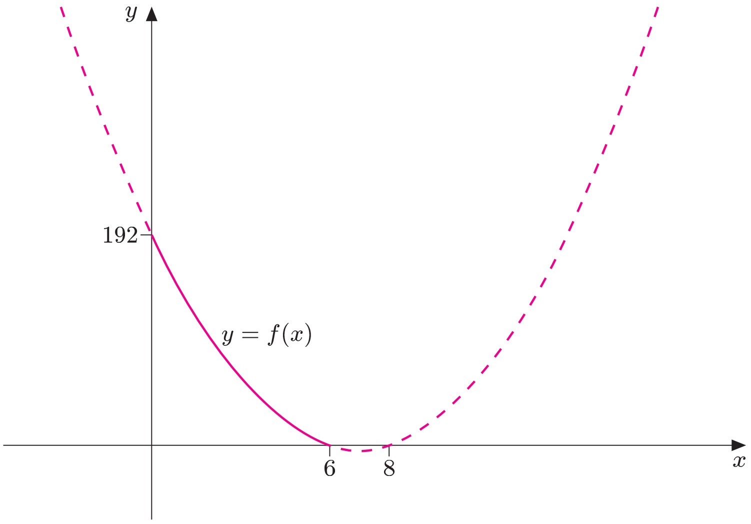


Figure 5.1 Graph of the function $f(x) = 4x^2 - 56x + 192$ (x in $[0,6]$)

The modified exhibition hall problem

Suppose now that you have to find the width of the border in the exhibition hall that results in the clear space having area 144. This problem leads to the equation

$$4x^2 - 56x + 192 = 144; \tag{5.3}$$

that is, $f(x) = 144$, where f is the function given in equation (5.2).

The packing case problem

The following problem is a three-dimensional version of the exhibition hall problem, with area replaced by volume.

One of the six sides is the lid.

The Packing Case Problem

A packing case is lined with polystyrene in such a way that each of the six sides has lining of equal thickness, and the resulting cavity is one third of the volume of the box. The dimensions of the box are 3 metres by 1 metre by 1 metre. What is the thickness of the lining?

We apply the usual approach to solving this problem: we introduce variables, state their ranges, find a formula for the volume of the cavity, and then use this formula to write down an equation and a relevant function.

Figure 5.2 shows the box with the polystyrene lining and the resulting cavity. Here we have called the volume of the cavity V, and the thickness of the polystyrene lining x. Clearly, x must lie in the interval $[0, 0.5]$.

Figure 5.2 represents a horizontal cross-section through the middle of the packing case.

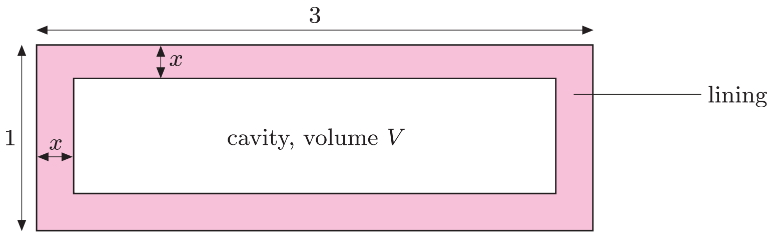


Figure 5.2 Variables for the packing case problem

The cavity has dimensions $3 - 2x$, $1 - 2x$ and $1 - 2x$, so its volume V is

$$\begin{aligned} V &= (3 - 2x)(1 - 2x)^2 \\ &= (3 - 2x)(1 - 4x + 4x^2) \\ &= 3 - 14x + 20x^2 - 8x^3. \end{aligned}$$

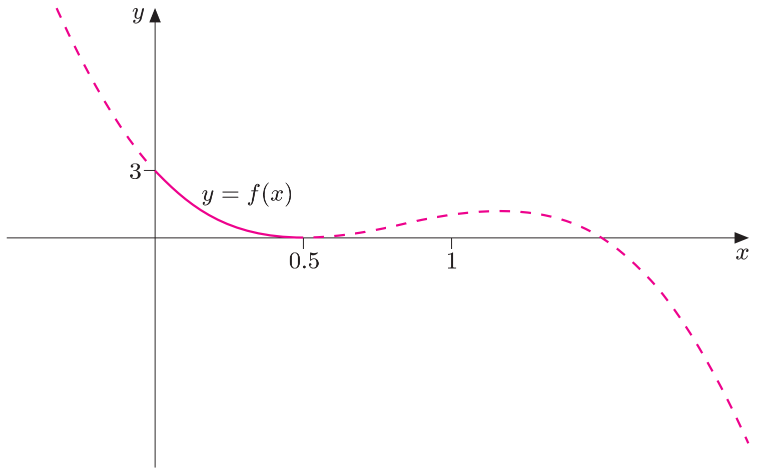
The problem is to find the value of x such that the volume V is one third of the volume of the box; that is, $V = \frac{1}{3} \times 3 \times 1^2 = 1$. Therefore we have to solve the cubic equation

$$3 - 14x + 20x^2 - 8x^3 = 1. \tag{5.4}$$

We represent the volume of the cavity using the cubic function

$$f(x) = 3 - 14x + 20x^2 - 8x^3 \quad (x \text{ in } [0, 0.5]). \tag{5.5}$$

Figure 5.3, on the next page, shows the graph of $y = 3 - 14x + 20x^2 - 8x^3$. The solid part is the graph of $y = f(x)$.



The graph of a cubic function may meet the x-axis once, twice (as here) or three times.

Figure 5.3 Graph of the function $f(x) = 3 - 14x + 20x^2 - 8x^3$ (x in $[0, 0.5]$)

5.2 Functions, graphs and solutions

In the first activity you are asked to define, in Mathcad, the function f given in equation (5.2), and then to find various values of this function.

Activity 5.1 Finding values of a function

Open Mathcad file **121A3-01 Functions graphs and solutions**. Read pages 2 and 3 of the worksheet, and carry out Task 1.

Solutions and some comments are provided in the Mathcad worksheet.

In the next activity you will plot the graphs of $y = f(x)$, where f is the function in equation (5.2), and of the line $y = 96$, over the interval $[0, 6]$. The solution to the exhibition hall problem is the value of x at the point where these two graphs meet. Mathcad's graph zoom facility enables you to enlarge a portion of the plot in order to read off the coordinates of this point with reasonable accuracy.

Activity 5.2 Using a graph to solve an equation

Read pages 4 and 5 of the worksheet, and carry out Tasks 2 and 3. These show you how to use Mathcad's graph zoom facility to find the solution to the exhibition hall problem.

You should still be working with Mathcad file 121A3-01.

Comments and a solution are provided in the Mathcad worksheet.

Mathcad notes

- ◇ The technique used to plot the line $y = 96$ can also be used to plot the x-axis, the line $y = 0$, on a Mathcad graph. The y-axis can be drawn in a similar manner, by plotting the line $x = 0$. Another method for adding axes is to format the graph and use 'Show markers'. See *A Guide to Mathcad* for more details about both of these methods.
- ◇ When plotting two or more y-axis expressions against the same x-axis expression, you need to enter the x-axis expression only once. So the graphs in the worksheet could be obtained by typing just `x` in the x-axis placeholder and `f(x),96` in the y-axis placeholder.

The final page of the worksheet from file 121A3-01 is a graphical solution template. The word 'template' indicates that this part of the worksheet has been prepared so that you can use it to solve a variety of problems by making only a small number of changes. This reduces work, but it has the disadvantage that the notation used in the template may not be the same as that used in any given problem.

Activity 5.3 Using the graphical solution template

You should still be working with Mathcad file 121A3-01.

Read page 6 of the worksheet. This contains the graphical solution template, set up for the exhibition hall problem. Task 4 is to use this template to solve the modified exhibition hall problem and the packing case problem.

(a) Use the template to solve the modified exhibition hall problem, given by equation (5.3). To do this, alter the target value to read

$$A := 144,$$

and then use Mathcad's graph zoom facility.

You can use the key sequence

```
3-14*x+20*x∧2
[Space]-8*x∧3
```

to create the right-hand side.

(b) Now use the template to solve the packing case problem, given by equation (5.4). Set up the problem by altering the function to read

$$f(x) := 3 - 14x + 20x^2 - 8x^3,$$

the target value to read

$$A := 1,$$

and the interval endpoints to read

$$X1 := 0 \quad \text{and} \quad X2 := 0.5.$$

Then use Mathcad's graph zoom facility.

Solutions are given on page 35.

Comment

◇ Remember that in part (b) the target value A represents *volume*.

◇ In the definition of the range variable x, the step size used is 0.01. This is the horizontal displacement between the points plotted, so it is not worth zooming in to give an x-interval narrower than 0.01. If you wish to obtain a more accurate solution graphically, then you should first reduce the step size appropriately.

The step size also affects the accuracy of values obtained using the graph trace tool.

Now close Mathcad file 121A3-01.

5.3 Solve blocks

In Subsection 5.2 you solved the exhibition hall problem, the modified exhibition hall problem and the packing case problem, by solving an appropriate equation in each case, using the graph of a relevant function. The next activity introduces the Mathcad 'solve block', and uses it to solve these equations directly. Once again the worksheet includes a template, set up so that you can apply the solve block to a variety of equations.

Activity 5.4 Using the solve block template

Open Mathcad file **121A3-02 The Mathcad solve block**. Read page 2 of the worksheet, which explains how to create a solve block.

Now read page 3 of the worksheet. This contains the solve block template, set up to solve the exhibition hall problem. Check that the solution given is the one you would expect. Task 1 is to use this template to solve the modified exhibition hall problem and the packing case problem.

(a) Use the template to solve the modified exhibition hall problem. To do this, alter the target value to read

$$A := 144.$$

(b) Now use the template to solve the packing case problem. Set up the problem by altering the function to read

$$f(x) := 3 - 14x + 20x^2 - 8x^3,$$

the target value to read

$$A := 1,$$

and the interval endpoints to read

$$X1 := 0 \quad \text{and} \quad X2 := 0.5.$$

Then enter a 'guess' for the value of '$x :=$' in the interval $(0, 0.5)$.

Solutions are given on page 35.

Comment

◇ The Mathcad solve block uses an 'iterative' numerical method to solve the equation; that is, starting with the value entered for the 'guess', it systematically refines guesses at a very fast rate to give an approximate solution, usually accurate to at least two decimal places. The initial guess may affect the solution obtained. Given the correct interval constraints, and a value for the guess within this interval, Mathcad will usually find an appropriate solution to the equation. If Mathcad cannot find a solution, then the expression '$\mathit{Find}(x) =$' appears in red, with the error message 'This variable is undefined.'.

Clicking on the red expression reveals the error message.

◇ Notice that the solve block uses *three* different forms of the 'equals' symbol, each with a different role.

The symbol ':=' is used to *define* (assign a value to) the variable on its left. For example,

$$x := 1$$

assigns the value 1 to the variable x.

The symbol '$\boldsymbol{=}$' is used to *equate* the expressions on either side of it. For example,

The bars of this equals sign are thicker than usual.

$$4x^2 - 56x + 192 \boldsymbol{=} 96$$

equates the expression $4x^2 - 56x + 192$ to the value 96.

The usual symbol '=' is used to *evaluate* and *display* the expression to its left. For example,

$$\mathit{Find}(x) = 0.917.$$

Now close Mathcad file 121A3-02.

5.4 Solving equations symbolically

The solve block is not the only way that Mathcad can solve equations. Some equations can be solved symbolically, that is, algebraically. This technique is used in the next two activities to solve the three problems in Subsection 5.1.

Activity 5.5 Solving quadratic equations

Open Mathcad file **121A3-03 Solving equations symbolically**.

(a) Read page 2 of the worksheet, and carry out Task 1, which uses the symbolic keyword 'solve' to solve the exhibition hall problem algebraically.

(b) Read page 3 of the worksheet, and carry out Task 2, which first uses the symbolic keyword 'solve' to solve the modified exhibition hall problem, and then shows how to obtain numerical rather than algebraic solutions.

Comments and solutions are provided in the worksheet, and further comments are below.

Comment

◇ Notice that the two solutions of the quadratic equation are displayed as a finite sequence in a column.

◇ There are no constraints on the solutions here, so the symbolic 'solve' gives every solution that it can find by using the quadratic equation formula. In each of parts (a) and (b), only one of the two 'solutions' found lies within the domain $[0, 6]$ of the function f given in equation (5.2).

◇ Some quadratic equations have no real solutions – that is, no solutions which are real numbers. In such cases, the symbolic 'solve' gives two solutions which are *complex numbers*. These involve i, a symbol with the property that $i^2 = -1$.

Complex numbers are studied in MS221.

In Activity 5.5, you obtained *exact* solutions, such as $7 + \sqrt{37}$ and $7 - \sqrt{37}$, to various quadratic equations. You also saw how to obtain the corresponding decimal solutions. The next activity asks you to apply a template for solving equations symbolically, which can be applied to quadratic or to more general equations.

Activity 5.6 Using the symbolic solution template

You should still be working with Mathcad file 121A3-03.

Look at page 4 of the worksheet, which contains a symbolic solution template. Use this template to solve the packing case problem, as follows.

(a) Set up the problem by altering the function to read

$$f(x) := 3 - 14x + 20x^2 - 8x^3$$

and the target value to be $A := 1$. The template then gives exact solutions to the equation $f(x) = 1$.

(b) Change the target value to $A := 1.0$ (that is, add a decimal point). Hence find, as a decimal, the solution to the packing case problem.

Comment

- ◇ The cubic equation $f(x) = 1$ here has the exact solutions 1, $\frac{3}{4} + \frac{1}{4}\sqrt{5}$, $\frac{3}{4} - \frac{1}{4}\sqrt{5}$. Mathcad's symbolic 'solve' can solve any cubic equation to give three exact solutions (two of which may be complex numbers). In general, however, the formulas for the solutions of a cubic equation are very long and can fill more than a page! You can avoid them appearing on this template in other cases, by making sure that at least one number in any cubic equation includes a decimal point, so that the solutions are displayed as decimals.
- ◇ The numerical solutions of the equation $f(x) = 1$ are 1, 1.3090... and 0.1909... . Only one of these lies in the domain $[0, 0.5]$ that was specified, in equation (5.5), for the packing case function. Hence the solution of that problem is 0.191 to 3 decimal places.

Now close Mathcad file 121A3-03.

You have now used three Mathcad techniques to solve various equations. As a general rule:

- ◇ a graphical solution can be used to solve *any* equation, but it is limited by the accuracy to which solutions can be found;
- ◇ the solve block can be used to solve *most* equations, including polynomial equations;
- ◇ the symbolic 'solve' can be used to solve many polynomial equations (and some others), but it will not always succeed.

For any given equation, a combination of these techniques may be used. For example, it is often useful to think about the approximate location of solutions graphically before finding their values numerically or symbolically.

5.5 A general graph plotter

The final activity uses a general graph plotter, which enables you to plot the graphs of a wide variety of functions. For example, you can use it to plot graphs of all the functions in the main text. You can also use it to solve two new problems: the orienteer's problem and the forester's problem. These are somewhat different to the problems solved earlier. In each case, a function f is given whose domain is an interval, and you are required to find the value of x in that interval for which $f(x)$ takes the *least* value.

The orienteer's problem

The sport of orienteering involves running or walking through the countryside, navigating between successive checkpoints. Orienteers are often faced with the following type of problem.

Here kph means ‘kilometres per hour’.

The Orienteer’s Problem

An orienteer, who can run at 16 kph on a straight path and at 8 kph in a forest, is at point A in the forest, 1 km away from the nearest point O on the path, and wishes to reach a point C on the path, which is 2 km from O, in the shortest time. At which point B should the orienteer aim to join the path?

All distances used here are in kilometres.

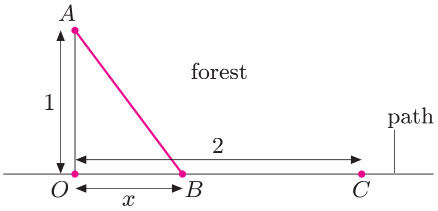


Figure 5.4 The orienteer’s problem

The variable x for the distance OB is introduced in Figure 5.4. The triangle AOB is a right-angled triangle, so AB is $\sqrt{1+x^2}$. Also, BC is $2-x$. Therefore the total time T taken (in hours) to run from A to C is

$$\begin{aligned} T &= \frac{\sqrt{1+x^2}}{8} + \frac{2-x}{16} \\ &= 0.125\sqrt{1+x^2} + 0.0625(2-x). \end{aligned}$$

We define the function f so that it represents the total time taken:

$$f(x) = 0.125\sqrt{1+x^2} + 0.0625(2-x) \quad (x \text{ in } [0,2]).$$

The orienteer’s problem is to find the distance x that minimises the time taken to run from A to C. This involves finding the value of x in the interval $[0,2]$ that gives the least value of $f(x)$. The graph of $y = f(x)$ is shown in Figure 5.5, and it can be seen from this that the required value of x is slightly more than 0.5 km.

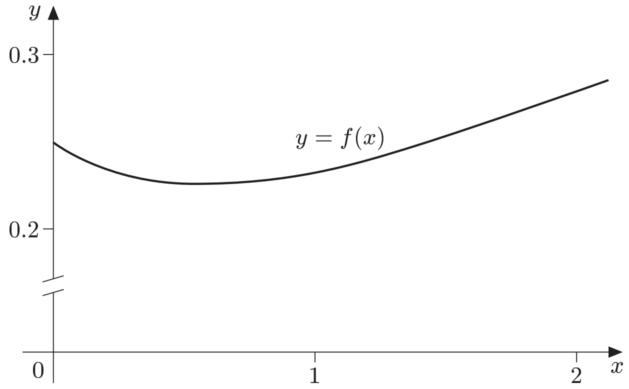


Figure 5.5 Graph of the function for the orienteer’s problem

The forester's problem

A forester wishes to remove a row of felled trees in a forest, using a stationary tractor and a winch with a 100-metre cable; see Figure 5.6. Those trees nearest to the tractor are removed first. It is assumed that the trees are of similar size and are distributed evenly along the row.

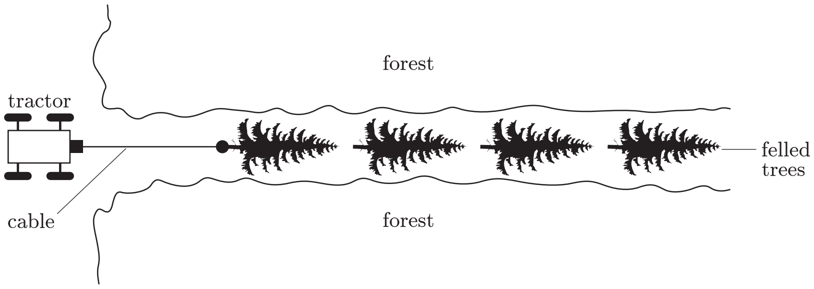


Figure 5.6 Removing a row of felled trees

The Forester's Problem
The cost (in pence) of removing a row, x metres long, of felled trees, where $0 \le x \le 100$, is modelled as

$$600 + 6x + 0.9x^2. \qquad (5.6)$$

For which value of x is the average cost per metre the least?

The cost formula in this model comprises three parts:

- the amount of 600 pence arises from fixed costs, independent of the length of the row x, such as the cost of preparing the tractor;
- the amount of $6x$ pence arises from costs which are the same for each tree removed, such as the time taken to attach the cable;
- the amount of $0.9x^2$ pence arises from costs which increase the further a tree is along the row, such as the time to carry the cable and the time to winch the tree to the tractor.

If it is known for which value of x the average cost per metre is least, then this knowledge might influence the possible locations of winching points in the forest. The above model is likely to be a great oversimplification, however.

The average cost per metre is

$$\frac{600 + 6x + 0.9x^2}{x} = \frac{600}{x} + 6 + 0.9x,$$

so we introduce the function

$$f(x) = \frac{600}{x} + 6 + 0.9x \quad (x \text{ in } (0, 100]).$$

The forester's problem is to find the number x that minimises the average cost per metre. This involves finding the value of x in the interval $(0, 100]$ that gives the least value of $f(x)$. The graph of $y = f(x)$ is shown in Figure 5.7, and it can be seen from this that the required value of x is approximately 30.

The value 0 is excluded from the domain of f since the average cost per metre does not make sense for this value of x.

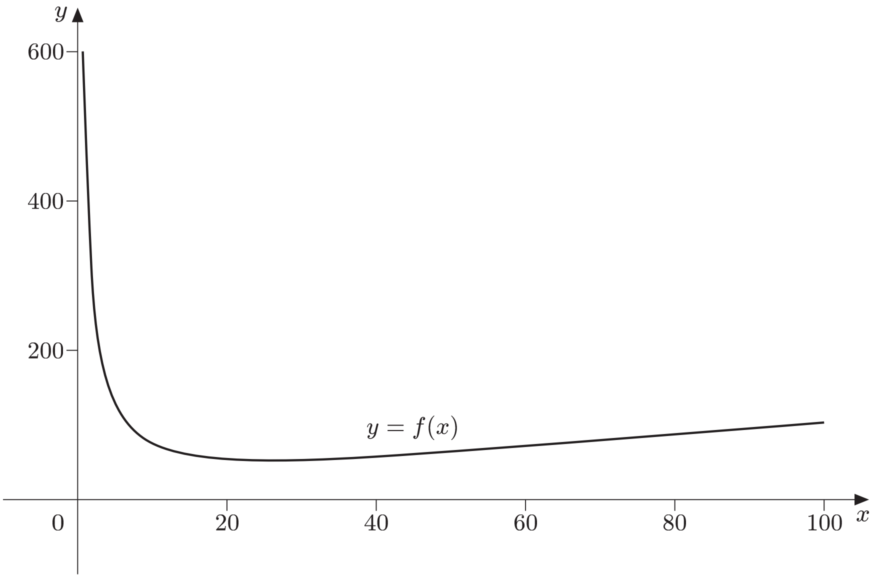


Figure 5.7 Graph of the function for the forester's problem

Activity 5.7 Using the Mathcad graph plotter

Open Mathcad file **121A3-04 Mathcad graph plotter**. The worksheet is set up to plot the function

$$f(x) = 0.125\sqrt{1 + x^2} + 0.0625(2 - x) \quad (0 \le x \le 2)$$

from the orienteer's problem. Task 1 is to use the graph plotter to solve the orienteer's problem and the forester's problem.

(a) Use the graph plotter to solve the orienteer's problem. To do this, use the graph zoom facility to zoom in on the portion of the plot where $f(x)$ takes the least value, and estimate the corresponding value of x by eye or with the graph trace tool.

(b) Now use the graph plotter to solve the forester's problem. Set up the problem by altering the function to read

$$f(x) := \frac{600}{x} + 6 + 0.9x,$$

and altering the interval endpoints to read

$$X1 := 1 \quad \text{and} \quad X2 := 100.$$

Then zoom in on the portion of the plot where $f(x)$ takes the least value, and estimate the corresponding value of x by eye or with the graph trace tool.

Solutions are given on page 35.

You can use the key sequence

```
600/x[Space]+6+0.9*x
```

to create the right-hand side.

The point 0 is not in the domain of f, so we take $X1$ slightly to the right of this.

Comment

- ◇ After you alter the function in part (b), Mathcad gives a graph which appears to have a value close to 3×10^5 at about $x = 0$. In fact, Mathcad has omitted to plot the point $(0, f(0))$, because $f(0)$ is not defined. The first point plotted is $(0.002, f(0.002))$, where $f(0.002) \simeq 3 \times 10^5$ (see the second Mathcad note below). The graph looks like that in Figure 5.7 once $X1$ is set to the new value 1 and $X2$ to 100.
- ◇ Finding the least (or the largest) value of a function graphically may involve inspecting a part of the graph which is very flat. It can be tricky to judge accurately where the least value occurs – do your best! Inspection of the y-values provided by the graph trace tool helps here.
- ◇ You are encouraged to try using the graph plotter to plot graphs of the various functions introduced in the main text. The expressions for several of these functions can be input either by using buttons on the 'Calculator' toolbar or by typing in directly. Thus there are buttons for both $|x|$ and $\sqrt{x}$ on the toolbar, but these can also be typed in directly as `|x` (the vertical bar is obtained from `[Shift]\`) and as `\x`, respectively. The function 2^x can be obtained either by typing `2∧x` or by using the 'x^y' toolbar button. Some standard functions cannot be obtained from toolbar buttons; for example, you need to type `acos(x)` in order to obtain $\arccos x$.

These alternatives were described for the functions sin and cos in the Mathcad notes for Activity 5.4 of Chapter A2, on page 18.

Mathcad notes

- ◇ The graph range is defined here as

$$x := X1, X1 + \frac{X2 - X1}{1000} \,..\, X2;$$

  that is, x ranges from $X1$ to $X2$ in steps of size $(X2 - X1)/1000$. This gives a way of plotting 1000 points, whatever values are chosen for the endpoints $X1$ and $X2$. As in Activity 5.3, this step size affects how far it is worth zooming in on the graph.

For example, in the forester's problem, the step size is

$$(100 - 1)/1000 = 0.099.$$

- ◇ If you try to plot the graph of a function f using a range variable x, and f is not defined at one of the values of x, then Mathcad avoids the problem by omitting any such value from the range.

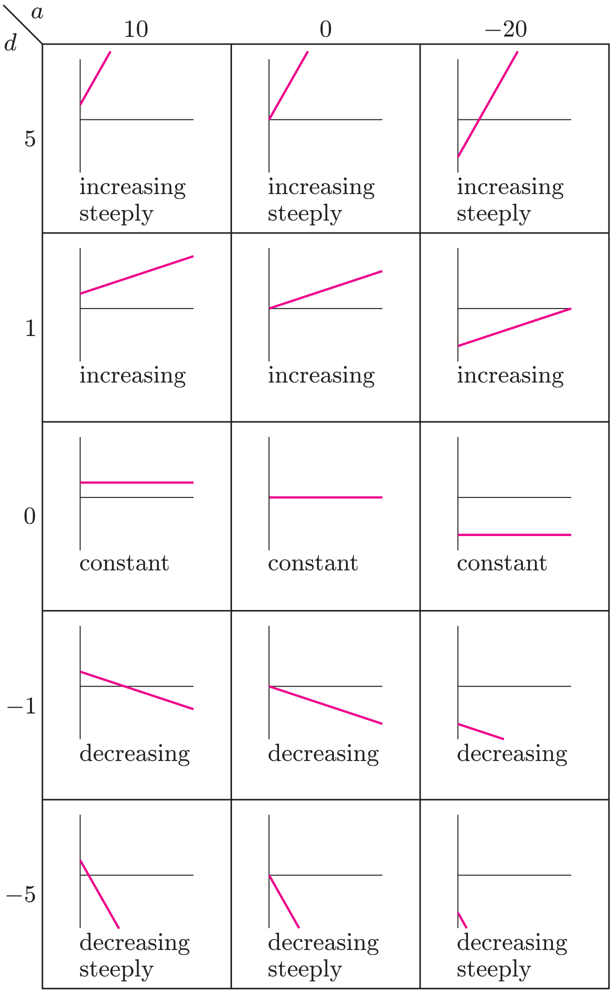
Now close Mathcad file 121A3-04.

Solutions to Activities

Chapter A1

Solution 6.2

For completeness, all the entries in the table have been included.



d \ a	10	0	−20
5	increasing steeply	increasing steeply	increasing steeply
1	increasing	increasing	increasing
0	constant	constant	constant
−1	decreasing	decreasing	decreasing
−5	decreasing steeply	decreasing steeply	decreasing steeply

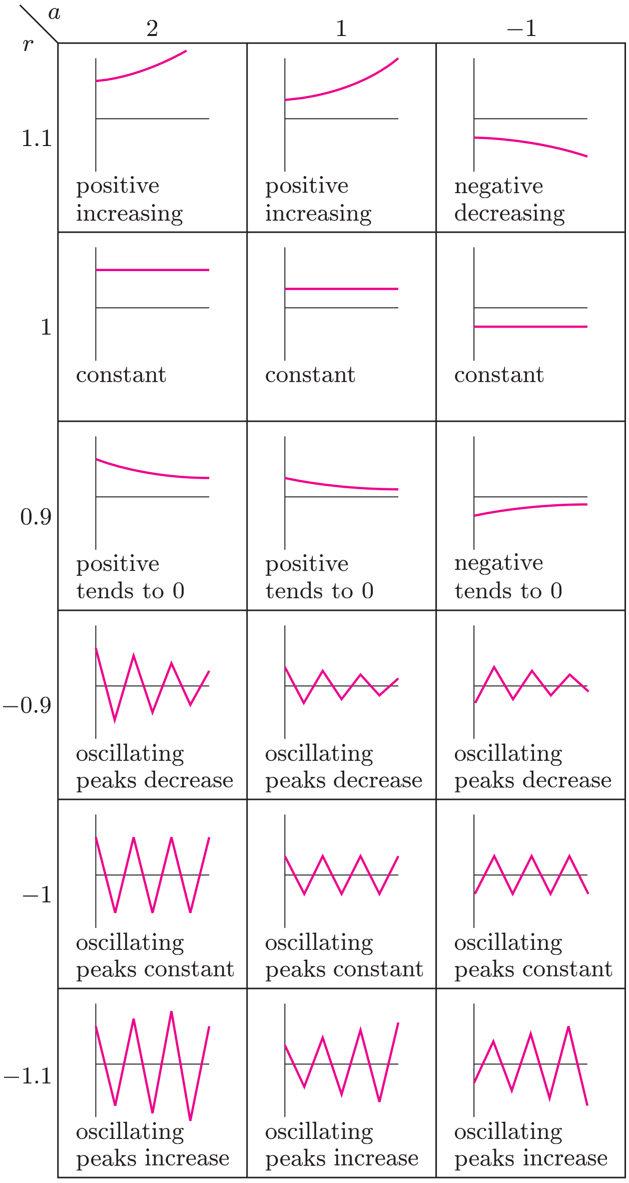
Figure S1.1 Sketch graphs of arithmetic sequences

(a) The parameter a is equal to the first term of the sequence; it affects where the graph starts.

(b) The parameter d determines the slope of the graph; that is, how steep the graph is (upwards or downwards).

Solution 6.3

The answers are included in the following table.



r \ a	2	1	−1
1.1	positive increasing	positive increasing	negative decreasing
1	constant	constant	constant
0.9	positive tends to 0	positive tends to 0	negative tends to 0
−0.9	oscillating peaks decrease	oscillating peaks decrease	oscillating peaks decrease
−1	oscillating peaks constant	oscillating peaks constant	oscillating peaks constant
−1.1	oscillating peaks increase	oscillating peaks increase	oscillating peaks increase

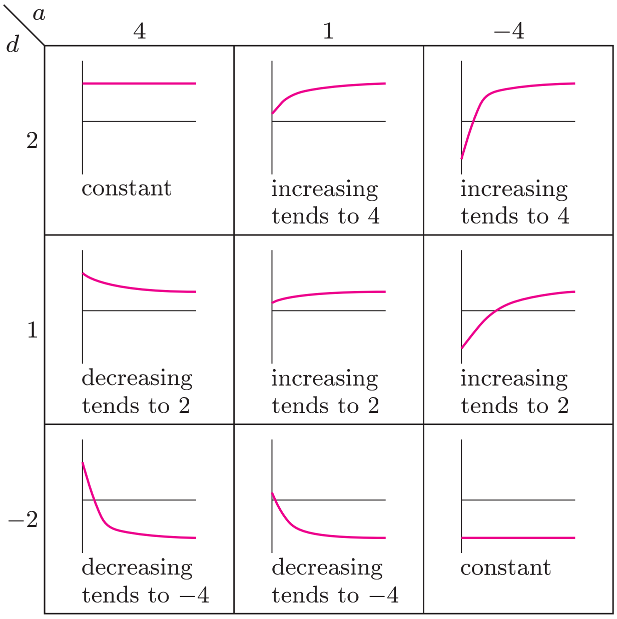
Figure S1.2 Sketch graphs of geometric sequences

(a) The parameter a is equal to the first term of the sequence; it affects where the graph starts and hence the orientation of the graph.

(b) The parameter r determines the *overall* shape of the graph – whether it is constant, tends to infinity, flattens out, or oscillates with constant, increasing or decreasing peaks.

Solution 6.4

The answers are included in the following table.



d \ a	4	1	-4
2	constant	increasing tends to 4	increasing tends to 4
1	decreasing tends to 2	increasing tends to 2	increasing tends to 2
-2	decreasing tends to -4	decreasing tends to -4	constant

Figure S1.3 Sketch graphs of linear recurrence sequences with $r = 0.5$

(a) Altering the parameter a changes the initial term, but has no effect on the value to which the sequence tends.

(b) Altering the parameter d changes the value to which the sequence tends, but has no effect on the initial term.

Chapter A3

Solution 5.3

(a) The solution to the modified exhibition hall problem is $x = 0.92$ (to 2 d.p.). This method is not very accurate, so any answer in the interval $[0.91, 0.93]$ is acceptable.

(b) The solution to the packing case problem is $x = 0.19$ (to 2 d.p.). This method is not very accurate, so any answer in the interval $[0.18, 0.20]$ is acceptable.

Solution 5.4

(a) The solve block gives the solution to the modified exhibition hall problem as $x = 0.917$, which is correct to 3 decimal places.

(b) The solve block gives the solution to the packing case problem as $x = 0.191$, which is correct to 3 decimal places.

Solution 5.7

(a) The solution to the orienteer's problem is $x = 0.58$ (to 2 d.p.). This method is not very accurate, so any answer in the interval $[0.57, 0.59]$ is acceptable. So in Figure 5.4 the point B should be 0.58 km from O.

(b) The solution to the forester's problem is $x = 26$, to the nearest integer. So the average cost per metre is least when the length of the row is 26 metres. This is quite accurate enough for practical purposes.

Computer Book B

Discrete Modelling

Guidance notes

This computer book contains those sections of the chapters in Block B which require you to use Mathcad. Each of these chapters contains instructions as to when you should first refer to particular material in this computer book, so you are advised not to work on the activities here until you have reached the appropriate points in the chapters.

In order to use this computer book, you will need the following Mathcad files.

Chapter B1

121B1-01	Logistic recurrence sequences
121B1-02	Overview of logistic recurrence sequences (Optional)

Chapter B2

121B2-01	Matrices in Mathcad
121B2-02	Exploring two subpopulations
121B2-03	Exploring three subpopulations

Chapter B3

There are no specified computer activities associated with Chapter B3.

Instructions for installing these files onto your computer's hard disk, and for opening them, are given in Chapter A0.

Activities based on software vary both in nature and in length. Sometimes the instructions for an activity appear only in the computer book; in other cases, instructions are given in the computer book and on screen. Feedback on an activity is sometimes provided on screen and sometimes given in the computer book.

For advice on how each computer session fits into suggested study patterns, refer to the Study guides in the chapters.

Chapter B1, Section 4
Logistic recurrence sequences on the computer

In this section, you will use the computer to investigate the behaviour in the long term of logistic recurrence sequences. There are two Mathcad files accompanying this section. The first file plots graphs of logistic recurrence sequences and enables the effect of altering parameters in the recurrence system to be investigated. The second file, which is *optional*, provides an overview of the types of long-term behaviour that can be obtained from a logistic recurrence sequence.

4.1 Investigating logistic recurrence sequences

See Chapter B1, Subsection 3.1.

The logistic model for population variation is described by the recurrence relation

$$P_{n+1} - P_n = rP_n\left(1 - \frac{P_n}{E}\right) \quad (n = 0, 1, 2, \ldots),$$

where P_n is the population size n years after some chosen starting time. The initial population size is P_0, and the other parameters are E, the equilibrium population level, and r, the proportionate growth rate at small population sizes.

In the first activity you will use the computer to investigate the effect of altering the value of P_0 on the long-term behaviour of the recurrence sequence. Then, in Activities 4.2 and 4.3, you will investigate the effect of altering the parameter r.

See Chapter B1, Subsection 3.3.

As you may recall, the parameter E is just a scaling factor, so no new types of long-term behaviour arise from altering the value of E.

Activity 4.1 Altering the initial population size

Remember to make your own working copy of the file.

This population was introduced in Example 3.1 of Chapter B1.

Recall that a scroll bar will appear when you click on a value in the table.

Open Mathcad file **121B1-01 Logistic recurrence sequences**. Page 1 introduces the worksheet. Look at page 2, where the task starts with the logistic model for the population of barnacle geese at Caerlaverock, Dumfries. The parameters for this model are $P_0 = 3200$, $r = 0.2$ and $E = 13\,300$. A table of values of terms of the sequence is given on this page, along with a fixed-scale graph that illustrates the sequence. Only terms of the sequence with values between 0 and 20 000 (shown as 2×10^4) can be shown on this scale.

Investigate the effect on the sequence of changing the value of the initial population size P_0, while keeping the parameters r and E constant. Set $N = 50$, so that longer-term behaviour can be observed. Use in turn the following values for P_0:

$$20\,000,\ 10\,000,\ 5000,\ 1000,\ 100.$$

Describe the long-term behaviour of these sequences, recording your observations in words and with small sketches. For example,

'the sequence is increasing and tends to E'

and Figure 4.2, when taken together, adequately describe the long-term behaviour of the sequence with $P_0 = 3200$.

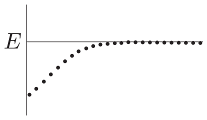


Figure 4.2 Small sketch to describe the long-term behaviour of a sequence

Do these sequences all tend to the same value?

Solutions are given on page 55.

Comment

The long-term behaviour of the sequence is *not* affected by the initial population size P_0, provided that P_0 lies between 0 and $E(1+1/r)$. For any such value of P_0, the sequence tends to E.

See Chapter B1, Subsection 3.3.

For the logistic recurrence relation with $r = 0.2$, we can say the following.

◇ For an initial population size P_0 below E but greater than 0, in this case $0 < P_0 < 13\,300$, the population increases and approaches the equilibrium level, $E = 13\,300$. The population then remains effectively constant at E.

◇ For an initial population P_0 above E but below $E(1+1/r)$, in this case $13\,300 < P_0 < 79\,800$, the population decreases initially and approaches the equilibrium level, $E = 13\,300$. Again, the population then remains effectively constant at E.

◇ If the initial population size P_0 is equal to the equilibrium level E, then the population remains constant at this value.

◇ If the initial population size P_0 is equal to $E(1+1/r)$, in this case $P_0 = 79\,800$, then $P_n = 0$ for all $n \geq 1$; that is, all subsequent terms of the sequence are zero. If the initial population size P_0 is greater than $E(1+1/r)$, then the next term of the sequence is negative. Therefore, the model predicts that for initial population sizes of $E(1+1/r)$ or more, the population becomes extinct within a year.

In fact, if P_0 is above E but below $E/r = 66\,500$, then the sequence decreases throughout. If P_0 is between E/r and $E(1+1/r)$, then the second term of the sequence is less than E, and thereafter the sequence behaves as for the case with $0 < P_0 < E$.

You may like to try in turn the values $P_0 = 79\,800$, $P_0 = 79\,810$ and $P_0 = 79\,790$. The table shows the detailed results in each case. In order to see all non-negative terms of these sequences on the graph, you will need to rescale the vertical axis, by setting $Y2 = 80\,000$ (say). For the case $P_0 = 79\,790$, the sequence shows a sharp drop in the first year, but the population then recovers and tends to $E = 13\,300$.

This is an extreme example of the type of behaviour described above for P_0 between E/r and $E(1+1/r)$.

Notice that the values in the table are given to 3 decimal places. When dealing with a population, as here, we round these values to the nearest whole number. This would not be necessary, however, if we were modelling a very large population and counting in millions.

Mathcad notes

Remember that Mathcad notes are *optional*.

◇ The range variable n used to calculate the terms of the sequence is defined as $n := 0, 1 \,..\, N-1$. This ensures that the subsequent definitions of the subscripted variables P_0 and P_{n+1}, given by

$$P_0 := 3200, \quad P_{n+1} := P_n\left[1 + r\left(1 - \frac{P_n}{E}\right)\right],$$

together define *all* the terms $P_0, P_1, \ldots, P_N$ and *no others*. Mathcad carries out the definition for P_{n+1} here N times, once for each of the values $0, 1, \ldots, N-1$ in the range of n. As it does so, the subscript $n+1$ takes each of the values in the range $1, 2, \ldots, N$ in turn.

◇ Mathcad variables can be defined more than once, so they can take different values at different places in the worksheet. For example, the range variable n is defined as $n := 0, 1 \,..\, N-1$ at the top of page 2 (to calculate the terms of the sequence) and as $n := 0, 1 \,..\, N$ lower down the page (to plot them on the graph).

◇ The vertical scale on the graph is fixed from $Y1$ to $Y2$. This has been done by defining two variables $Y1$ and $Y2$ above the graph, then entering $Y1$ and $Y2$ into the axis limit placeholders. The graph can thus be rescaled by changing the value of $Y1$ or $Y2$, without the need to edit the graph itself. (The horizontal scale has also been fixed, from 0 to N.)

You are now asked to use the computer to investigate the effect of altering the parameter r. Taking values of r greater than 3 gives sequences whose terms quickly become very large, so we concentrate on values in the interval $(0, 3]$.

Interval notation was introduced in Chapter A3, Subsection 1.1.

Activity 4.2 Altering the parameter r

You should still be working with Mathcad file 121B1-01. If you previously altered the value of $Y1$ or $Y2$, then these should now be reset to $Y1 = 0$ and $Y2 = 20\,000$.

Investigate the effect of altering the parameter r on the long-term behaviour of the logistic recurrence sequence. Set $P_0 = 5000$, $N = 50$ and leave $E = 13\,300$.

Use in turn the following values for r:

$0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3.$

In some cases (in particular, for $r = 2$, 2.5 and 3), you may find it helpful to increase the value of N to 100, 500 or perhaps even more.

Record your observations in words and with small sketches.

Solutions are given on page 55.

Comment

Different values of r show markedly different types of behaviour.

◇ For $r = 0.5$ and $r = 1$, the sequence increases and tends to $E = 13\,300$. This is the same type of behaviour as displayed by the sequence in Activity 4.1, where $P_0 = 5000$, $E = 13\,300$ and $r = 0.2$.

In the case of $r = 2$, the convergence to E is very slow and can be seen convincingly only by choosing a large value for N.

◇ For $r = 1.5$ and $r = 2$, the terms of the sequence alternate between values above and below $E = 13\,300$, gradually approaching E. Hence the sequence tends to E.

◇ For $r = 2.5$, the terms of the sequence again alternate between values above and below $E = 13\,300$. In this case, however, the terms *do not* approach E. In the long run, each term is close to one of *four* values: 9326, 16 292, 7128 and 15 398 (to the nearest integer). This type of behaviour, of repeatedly taking a number of values in order, is called **cycling**. Here four repeating values are involved, so we have a **4-cycle**.

For example, these are the rounded values of P_{91}, P_{92}, P_{93}, P_{94}, respectively. They repeat as P_{95}, P_{96}, P_{97}, P_{98}, and so on.

◇ For $r = 3$, another type of long-term behaviour is displayed. Again, terms of the sequence do not approach $E = 13\,300$, as can be seen clearly by increasing N to 500 or 1000. The sequence behaves in an unpredictable manner, taking values which seem to be random.

However, we can at least say that terms of the sequence lie in the range $0 < P_n < 20\,000$, although, of course, no term is actually equal to $E = 13\,300$. This type of apparently unstructured behaviour is referred to as **chaotic**.

In fact, it looks as if P_n is always less than 18 000.

The values for P_0 and E were chosen for convenience and clarity. Similar results would be found using other values of E and values of P_0 between 0 and $E(1 + 1/r)$.

The six values of r used in Activity 4.2 were chosen to be equally spaced in the range from 0 to 3. The next step in a systematic investigation of this type is to pick values of r in the intervals between those pairs of values from Activity 4.2 which gave different types of long-term behaviour. For example, the three values 1.25, 2.25 and 2.75 are sensible choices. We ask you to continue this investigation in Activity 4.3.

Activity 4.3 Using further values for the parameter r

Investigate further the effect of altering the parameter r on the long-term behaviour of the logistic recurrence sequence. Again set $P_0 = 5000$, $N = 50$ and $E = 13\,300$.

You should still be working with Mathcad file 121B1-01.

Use in turn the following values for r:

1.25, 2.25, 2.75.

Record your observations in words and with small sketches. (Remember that you may have to increase the value of N to observe the long-term behaviour.)

Solutions are given on page 56.

Comment

Only one new type of behaviour is shown by sequences with these values of r.

◇ For $r = 1.25$, the terms of the sequence alternate between values just above and just below $E = 13\,300$, and the sequence tends to E. This is the same type of long-term behaviour as displayed in Activity 4.2 by the sequences with $r = 1.5$ and $r = 2$.

◇ For $r = 2.25$, the terms of the sequence alternate between values above and below $E = 13\,300$. The sequence *does not* tend to E, but settles to one value (15 608) above E and one value (9515) below E. We describe this type of long-term behaviour as a **2-cycle**.

Both values here are rounded to the nearest integer.

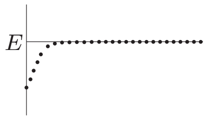
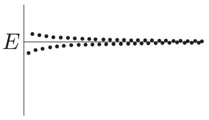
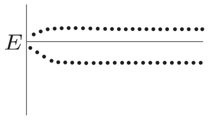
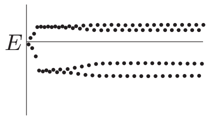
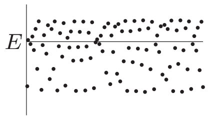
◇ For $r = 2.75$, the terms of the sequence take values which seem to be random between two bounds. The long-term behaviour is chaotic, as displayed in Activity 4.2 by the sequence with $r = 3$.

Now close Mathcad file 121B1-01.

On the basis of the cases that you examined using Mathcad in Activities 4.2 and 4.3, the table on page 42 can be drawn up to describe the long-term behaviour of the sequences.

Further investigation would be required to determine a fuller range of behaviour. It would clearly be possible to obtain more information by trying many more values of r, in an attempt to 'fill in the gaps'. In some cases this can give rapid answers. For example, it does not take much experimentation to establish with some confidence that the behaviour described in the table for $r = 0.5$ and $r = 1$ is shared by all values of r in the interval $(0, 1]$, whereas any value in the interval $(1, 2]$ gives the behaviour described for $r = 1.25$, 1.5 and 2.

However, rather than looking at a succession of individual cases, we really need some sort of 'overview' of what is going on. This is the topic of the next subsection.

r	Long-term behaviour of P_n	Sketch graph
0.5, 1	Tends to E, with values always just below E	
1.25, 1.5, 2	Tends to E, with values alternating between just above and below E	
2.25	2-cycle, with one value above E and one value below E	
2.5	4-cycle, with two values above E and two values below E	
2.75, 3	Chaotic variation between bounds	

4.2 An overview of long-term behaviour

Rather than look in turn at the sequences determined by particular values of r, it is possible to obtain an overview of the long-term behaviour of the sequences as the parameter r varies. Computers are most appropriately used for tasks that involve a relatively simple specification but much calculation, and this is very much a case in point.

The recurrence relation

This is derived in Chapter B1, Subsection 3.3.

$$x_{n+1} = x_n + rx_n(1 - x_n) \quad (n = 0, 1, 2, \ldots),$$

in which the parameter E does not appear, is used for simplicity. This is the logistic recurrence relation with $E = 1$ or, alternatively, the logistic recurrence relation rewritten using the substitution $x_n = P_n/E$.

In the previous two activities, the terms of each sequence were plotted from left to right, as a graph of P_n against n, giving a different graph for each value of the parameter r used. We now construct *one* plot that indicates the variety of types of long-term behaviour that can occur.

This overview is obtained as follows. A given starting value x_0 is defined. For each value of r in the specified range, the first 301 terms of the sequence, $x_0, x_1, \ldots, x_{300}$, are calculated, and then the first 200 terms, $x_0, x_1, \ldots, x_{199}$, are discarded. The remaining terms of the sequence are then plotted against the value of the parameter r. So, for each value of r, the graph displays each of the points (r, x_{200}), (r, x_{201}), ..., (r, x_{300}) as a small dot.

Figure 4.3 shows this type of overview as generated by Mathcad. The values of the parameter r used for this graph are $0, 0.002, 0.004, \ldots, 3$.

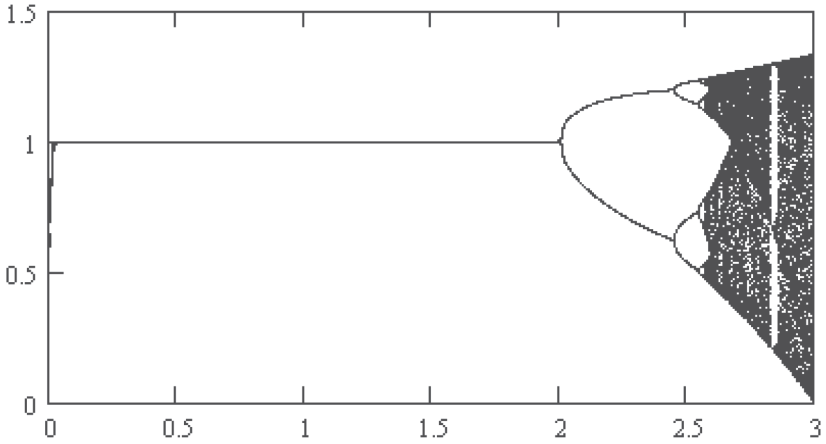


Figure 4.3 Overview of long-term behaviour of logistic recurrence sequences for $0 \leq r \leq 3$

To find the behaviour of the logistic recurrence sequence for a given value of r, you look along the horizontal axis to find the value of r, then look up the corresponding vertical direction at the values of the sequence plotted against this value of r. For example, look along the horizontal axis of the graph in Figure 4.3 for the value $r = 1.5$. Then look vertically to see which points, and in particular how many points, have been plotted against this value of r. In this case, there appears to be just one point plotted: $(1.5, 1)$. This indicates that (to the accuracy shown on the graph) all of the terms $x_{200}, x_{201}, \ldots, x_{300}$ of this sequence are equal to 1; that is, for this value of r, the sequence tends to the equilibrium population level $E = 1$, as expected from our previous investigations.

It can be seen from Figure 4.3 that the more complicated behaviour occurs for values of r greater than 2. Figure 4.4 shows another overview generated by Mathcad. For this graph, the values for the parameter r are $2.4, 2.4004, 2.4008, \ldots, 3$.

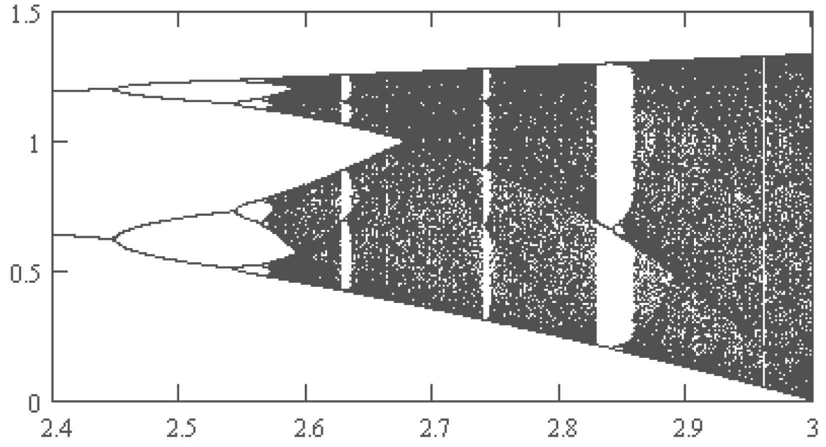


Figure 4.4 Overview of long-term behaviour of logistic recurrence sequences for $2.4 \leq r \leq 3$

Look along the horizontal axis of the graph in Figure 4.4 for the value $r = 2.5$, then look vertically to see how many points have been plotted against this value. In this case there appear to be four points plotted, two above and two below the value $E = 1$. This means that each of the terms $x_{200}, x_{201}, \ldots, x_{300}$ of this sequence takes one of these four values. Our findings from the previous investigations confirm that the long-term behaviour of the sequence with $r = 2.5$ is a 4-cycle.

In general, we can deduce that if there are a small number of points plotted on the vertical line for a given value of r, then the terms of the corresponding sequence approach a cycle with this number of values. For example, it looks as if there may be values of r between 2.5 and 2.6 where the long-term behaviour of the sequence is an 8-cycle.

For many values of r, a 'smear' of points is plotted in Figure 4.4. This indicates chaotic behaviour, where the terms $x_{200}, x_{201}, \ldots, x_{300}$ of the sequence take seemingly random values within two bounds. One of these bounds is above $E = 1$, and the other is below. Each of the bounds varies (smoothly) as r varies.

Note that the regime for chaos, beginning somewhere between $r = 2.5$ and $r = 2.6$, does *not* continue unbroken to $r = 3$. Indeed, there are values of r between 2.8 and 2.9 where it appears that the long-term behaviour of the sequence may be a 3-cycle or perhaps even a 6-cycle!

The following optional activity investigates further this use of Mathcad to provide an overview of the long-term behaviour of logistic recurrence sequences. You have seen that a relatively simple-looking recurrence relation generates sequences with remarkably rich patterns of behaviour. Other non-linear recurrence relations lead to families of sequences which behave in a similarly complicated way.

More explanation of why some of these patterns arise from such recurrence relations is given in MS221 Chapter B1.

Activity 4.4 An overview of long-term behaviour (Optional)

Open Mathcad file **121B1-02 Overview of logistic recurrence sequences**. Look at the graph plotted in this worksheet. This graph gives an overview of all the types of long-term behaviour that logistic recurrence sequences exhibit for values of the parameter r between 1.5 and 3. Bear in mind that Mathcad could take several seconds to perform the calculations and plot the graph. (Just how long it takes for the graph to be plotted depends on the speed of your computer.)

The graphs plotted in Figures 4.3 and 4.4 were generated with the variable V set to 1500. In this worksheet, V is set to 500 to reduce the time taken to produce data for the graph.

You may like to investigate the types of long-term behaviour displayed for a narrower range of values of r. For example, if you set $R1 = 2.8$ and leave $R2 = 3$, then this region of the graph will be expanded.

If you seek to change *both* of $R1$ and $R2$, then you will find that Mathcad starts to recalculate after the first change has been made. The first two notes in the Comment below explain how this can be avoided.

Can you find a value of r for which the long-term behaviour of the sequence appears to be a 3-cycle?

A solution is given on page 56.

Comment

◇ You can interrupt a Mathcad calculation by pressing `[Esc]`, the escape key, and then clicking 'OK' in the resulting option box. To resume calculation, press the `[F9]` function key, or go to the **Tools** menu, then choose **Calculate** and click on **Calculate Now**.

◇ By default, Mathcad operates in 'automatic calculation mode', but this can be inconvenient where more than one input change is to be made before recalculation is required. In order to switch to 'manual calculation mode' (which disables automatic calculation), select **Calculate** from the **Tools** menu, then click on either **Automatic Calculation** or the tick beside it. (When you do this, the tick mark beside **Automatic Calculation** in the menu disappears, and the word 'Auto' in the status bar, at the bottom right corner of the Mathcad window, is replaced by 'Calc F9'.)

Once in manual mode, you can calculate the results as and when you choose, either by selecting **Calculate** from the **Tools** menu and then clicking on **Calculate Now** or by pressing the `[F9]` function key.

In order to return to automatic mode, follow the same procedure, whereupon the tick mark against **Automatic Calculation** reappears.

More information about this technique, and details of how to interrupt and resume calculations, are provided in *A Guide to Mathcad*.

◇ If your computer is taking a long time to calculate all the terms of the sequences and to plot the graph, you may wish to speed up the process by reducing the amount of computation required. The parameter V determines the number of values of r for which the sequence is calculated, so by decreasing V, the sequence is calculated for fewer values of r, and the computation time is reduced. However, the resulting graph will appear sketchier as a consequence.

Mathcad notes

◇ The graph is obtained by plotting $x_{i,n}$ against r_i using the trace type 'points'. The subscripted variable $x_{i,n}$ is obtained in the usual way: either use the 'x_n' button on the 'Matrix' toolbar or type `[` (left square bracket), then separate the subscripts i and n with a comma. Thus you could obtain $x_{i,n}$ on the screen by typing `x[i,n`.

◇ A Mathcad graph can display approximately 490 000 individual points. If you attempt to plot a graph with more points than this, then an error may occur. (No graph is drawn and the graph box is highlighted in red – clicking on it reveals the error message 'Unable to plot this many points.'.)

Now close Mathcad file 121B1-02.

Chapter B2, Section 4
Computing with matrices

In this section, you will be working with matrices in Mathcad. The first of the three accompanying files shows how to create matrices and calculate with them in Mathcad. In the second and third files you will study further the long-term behaviour of subpopulations within matrix population models, as introduced in Section 3.

4.1 Matrices in Mathcad

See Chapter B2, Section 2.

Recall that an $m \times n$ matrix has m rows and n columns. In Activity 4.1 you will learn how to create, define and edit matrices in Mathcad.

Activity 4.1 Creating, defining and editing matrices

Open Mathcad file **121B2-01 Matrices in Mathcad**. Read page 2 of the worksheet, follow the instructions, and carry out Task 1.

See Chapter B2, Section 2.

In the main text, you investigated matrix multiplication, matrix addition and scalar multiplication of matrices. In the next activity, you will see how to use the computer to perform these operations.

Activity 4.2 Matrix arithmetic in Mathcad

You should still be working with Mathcad file 121B2-01.

Read page 3 of the worksheet, and carry out Task 2.

If you feel you need, or would like, more practice with multiplying matrices, then read the *optional* final page of the worksheet and perform the matrix multiplications given there.

Solutions are given on page 56.

Mathcad notes

- When multiplying two matrices you must enter the multiplication, either by clicking on the multiplication button '×' on the 'Calculator' toolbar or by typing *. In contrast, when multiplying a matrix on the left by a scalar (number), Mathcad will insert the multiplication if you omit to enter it. (In both cases, the multiplication appears initially as a small raised dot then, once entry is complete, this dot disappears from view.)
- Variables in Mathcad are case-sensitive; that is, upper-case (capital) letters and lower-case (small) letters are regarded as different.

When attempts are made to perform illegal operations on matrices, Mathcad responds with various error messages.

- Matrix addition, $\mathbf{A}+\mathbf{B}$, is defined only if the sizes of the two matrices $\mathbf{A}$ and $\mathbf{B}$ are equal, and matrix multiplication, $\mathbf{AB}$, is defined only if the number of columns of the matrix $\mathbf{A}$ is equal to the number of rows of the matrix $\mathbf{B}$. The Mathcad error message for an incorrectly formed matrix sum or product is 'These array dimensions do not match.'.
- Matrix powers, $\mathbf{A}^2$, $\mathbf{A}^3$, ..., are defined only if $\mathbf{A}$ is a square matrix. The Mathcad error message for trying to find powers in other cases is 'This matrix must be square.'.
- A curious exception to the previous comment is that Mathcad will 'calculate a power' of a vector (one-column matrix), being a vector of the same size each of whose elements is the given power of the corresponding original element. This is not a normally recognised operation on vectors.

Mathcad uses the word 'array' to describe a matrix or vector.

Now close Mathcad file 121B2-01.

You will study matrix arithmetic further in Chapter B2, Section 5. Two important concepts that you will learn about there are the *inverse* of a matrix and the *determinant* of a matrix. Mathcad can be used to find inverses and determinants of matrices, so we include the entry instructions here.

You may prefer to omit the remainder of this subsection for the moment, and return to it after studying Chapter B2, Section 5.

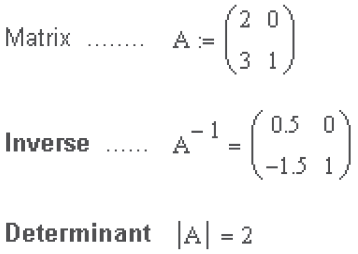
The inverse of a (square) matrix is denoted by the matrix raised to the power -1; for example, $\mathbf{A}^{-1}$ denotes the inverse of the square matrix $\mathbf{A}$. To find the inverse of a square matrix in Mathcad, either use the 'X^{-1}' button on the 'Matrix' toolbar or type `^-1`. For example, to display the inverse of the matrix $\mathbf{A}$, you could type `A^-1=`. (Note that the inverse of a square matrix does not always exist; where it does not, Mathcad responds with the error message 'This matrix is singular. Cannot compute its inverse.'.)

The inverse of a matrix is another matrix of the same size, whereas the determinant is a single number. Both are defined only for square matrices.

To calculate the determinant of a square matrix, either use the '$|X|$' button (available on both the 'Calculator' and 'Matrix' toolbars) or type `[Shift]\` (shift and backslash). For example, to display the determinant of the matrix $\mathbf{A}$, you could type `[Shift]\A=`.

The notations det $\mathbf{A}$, as defined in Chapter B2, Section 5, and $|\mathbf{A}|$ are both commonly used to denote the determinant of $\mathbf{A}$.

A Mathcad display of inverse and determinant, for a particular matrix $\mathbf{A}$, is as follows.



Matrix .... $A := \begin{pmatrix} 2 & 0 \\ 3 & 1 \end{pmatrix}$

Inverse ...... $A^{-1} = \begin{pmatrix} 0.5 & 0 \\ -1.5 & 1 \end{pmatrix}$

Determinant $|A| = 2$

It is also possible to find the inverse and determinant algebraically, using the buttons '$M^{-1} \to$' and '$|M| \to$' on the 'Symbolic' toolbar.

4.2 Matrix population models on the computer

See Chapter B2, Subsection 3.1.

In the main text, you investigated a population model of the UK with two subpopulations: juveniles (aged up to 15 years old) and adults. This model was based upon the sizes of these subpopulations in 1990 and on their birth and death rates. The matrix recurrence relation for this model is

Here J_n and A_n represent, respectively, the numbers of juveniles and adults at n years after the starting date in 1990.

$$\begin{pmatrix} J_{n+1} \\ A_{n+1} \end{pmatrix} = \begin{pmatrix} 0.9326 & 0.0172 \\ 0.0666 & 0.9864 \end{pmatrix} \begin{pmatrix} J_n \\ A_n \end{pmatrix}.$$

You looked at predictions of subpopulation sizes from the model for the years 1990–1997, that is, for $n = 0, 1, \ldots, 7$. This matrix model can be implemented on the computer, which calculates many terms quickly and plots graphs easily, so the long-term behaviour of the model can be observed directly. In the next activity, you will use Mathcad to investigate further the question 'Will there be a constant supply of juveniles entering the potential workforce, or will we have an ever-increasing proportion of adults?'.

Activity 4.3 The two-subpopulation matrix model

Open Mathcad file **121B2-02 Exploring two subpopulations**. Look at page 2 of the worksheet, where the task starts with the two-subpopulation model of the UK population. The parameter N, the number of subsequent years for which the model predicts, is set at 10. Mathcad therefore calculates the subpopulation sizes for the next 10 years, that is, for the years 1991–2000. These subpopulation sizes and the total population size are displayed in tables at the bottom of this page of the worksheet.

The values in the first 8 lines of these tables agree with the values given in Chapter B2, Subsection 3.1, for the years 1990–1997.

Look at page 3 of the worksheet. The first graph shows the sizes of the two subpopulations and the total population. The scale of this graph is quite large, but a slight increase in each of the subpopulations is discernible. The two graphs at the bottom of the page show the two subpopulations separately. Both of these graphs are clearly increasing; however, Mathcad automatically sets the axis limits for both graphs, so the scales are different. The number of juveniles increases by about half a million, whereas the number of adults increases by twice as much. The remaining graph on this page of the worksheet shows that the ratio of successive total populations, T_n/T_{n-1}, is decreasing. A value greater than one for this ratio indicates that the total population is increasing, a value less than one indicates that it is decreasing, and the value one indicates that it is constant year on year. A look at the scale shows that the decrease of the ratio T_n/T_{n-1} here is not significant: it is less than 0.000 02 over the ten years, and the ratio remains greater than one.

This ratio can be expressed as

$$\frac{T_n}{T_{n-1}} = 1 + \frac{T_n - T_{n-1}}{T_{n-1}},$$

that is, one plus the proportionate growth rate (as defined in Chapter B1, Subsection 3.1).

Now look at the graphs on page 4 of the worksheet. The top graph shows the proportions of juveniles and adults relative to the total population; these look fairly constant. The bottom two graphs, which use Mathcad's automatic scaling, look very different; these graphs show the proportion of juveniles increasing and the proportion of adults decreasing. However, a look at the scales reveals that these changes are in fact both less than 0.005 over the ten years.

(a) Return to page 2 of the worksheet, and set the value of N to 50. This changes the value of N throughout the worksheet.

Look at the graphs on page 3. What can you say about the subpopulation sizes and the ratio of successive total populations?

(b) Look at the graphs on page 4. Are the proportions of juveniles and adults eventually fairly constant? Try to answer the question 'Will there be a constant supply of juveniles entering the potential workforce, or will we have an ever-increasing proportion of adults?'.

Confirm your answers by changing N to 100.

Solutions are given on page 56.

Comment

◇ The graphs have been set up to show the total population traces as solid red lines, the adult traces as solid blue lines, and the juvenile traces as solid magenta (purple) lines. If you have any difficulty distinguishing between these colours (especially on the total population graph, where all three are plotted together), or wish to print these graphs on a non-colour printer, then you may like to format the graph to change the line style of some of these traces from solid to dotted or dashed.

◇ The long-term behaviour can be seen clearly when N is set to 100; this confirms the trends shown when N is set to 50. However, over such a long period, it is extremely likely that social changes will invalidate the model.

◇ The graphs in this worksheet show how important it is always to look at the scale before drawing any conclusions.

Mathcad notes

◇ The calculation range is $n := 0, 1 \mathinner{..} N-1$, and the subpopulations calculated at each stage are J_{n+1} and A_{n+1}. Hence the subpopulations $J_1, J_2, \ldots, J_N$ and $A_1, A_2, \ldots, A_N$ are calculated (J_0 and A_0 being specified beforehand). The graph range is $n := 0, 1 \mathinner{..} N$, so the subpopulations $J_0, J_1, \ldots, J_N$ and $A_0, A_1, \ldots, A_N$ are plotted.

◇ You may wonder why the graph range $n := 0, 1 \mathinner{..} N$ is not used to plot the ratio of successive total populations, T_n/T_{n-1}. The denominator, T_{n-1}, is not defined when $n = 0$ (there is no term T_{-1}), so we need to define a separate graph range variable k, taking only the values $1, 2, \ldots, N$, and then plot T_k/T_{k-1}.

◇ Titles have been added to the graphs in this worksheet. This is done by first clicking in the graph to select it, and then choosing **Graph ▶ X-Y Plot...** from the **Format** menu, to bring up the 'Formatting Currently Selected X-Y Plot' option box. The title is entered on the 'Labels' tab in the option box.

Alternatively, double-click in the graph to bring up the option box.

In the next activity you will investigate the effect of altering the initial size of the juvenile subpopulation on the long-term behaviour of the two-subpopulation model.

Activity 4.4 Altering the initial juvenile subpopulation size

You should still be working with Mathcad file 121B2-02.

On page 2 of the worksheet, set $N = 50$, so that the long-term behaviour can be seen.

(a) Use the following initial values for the subpopulation sizes:

$$\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{pmatrix} 5 \\ 46.49 \end{pmatrix}.$$

Use the tables on page 2 of the worksheet and the graphs on pages 3 and 4 to answer the following questions.

(i) Does the ratio of successive total populations tend to a limit? If so, what is this limit?

(ii) What happens to the numbers of juveniles and adults over the 50-year period of prediction?

(iii) What happens to the proportions of juveniles and adults over the 50-year period of prediction?

(b) Repeat part (a) with the following initial values for the subpopulation sizes:

$$\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{pmatrix} 20 \\ 46.49 \end{pmatrix}.$$

(c) Does the long-term behaviour of the model appear to depend on the initial values of the subpopulation sizes?

Solutions are given on page 57.

Comment

See Chapter A1, Subsection 6.3.

- ◇ In general, the initial values of the subpopulation sizes do not affect the long-term behaviour of the model. This conclusion parallels the results that you discovered in Block A for linear recurrence sequences. You may like to try more substantial changes to the initial subpopulation sizes. The ratio of successive total populations may increase or decrease in the early years, but it always seems to tend to 1.003 (to 3 decimal places), while the proportions of juveniles and adults seem to tend respectively to 0.2 and 0.8 (to 1 decimal place). However, in order to observe this long-term behaviour, the length of time N (years) over which the model is run may need to be increased in some cases to 100 or more.
- ◇ The accuracy of readings taken from these graphs varies, depending on what scale is chosen by the automatic scaling within Mathcad.

Now close Mathcad file 121B2-02.

Mathcad is an ideal tool for studying more complicated matrix population models. In the next activity, you will investigate a question posed in the main text: 'Who will support us in our old age?'

Activity 4.5 The three-subpopulation matrix model

Open Mathcad file **121B2-03 Exploring three subpopulations**. The layout of this worksheet is very similar to that of the previous one, with tables giving the values of the subpopulations on page 2, graphs of the subpopulation sizes and the ratio of successive total populations on page 3, and the proportions of the subpopulations on page 4. The obvious difference is that there are three subpopulations included in this model: juveniles (aged up to 15 years old), workers (aged between 15 and 65 years old) and the elderly.

Look at the tables and graphs in the worksheet. At first sight, the outlook predicted by the model is that the number of elderly grows, while the number of workers falls.

Change the number of years for which the model predicts (on page 2 of the worksheet) to $N = 50$.

(a) (i) Does the ratio of successive total populations tend to a limit? If so, what is this limit and what is the consequence for the population as a whole?

(ii) What happens to the proportions of juveniles, workers and elderly over the 50-year period of prediction?

(b) Confirm your observations from part (a) by changing N to 100.

(c) Try to answer the question 'Who will support us in our old age?', by looking at the ratio of the elderly to the workers who will support them.

Solutions are given on page 57.

Comment

You may have been surprised to see that the long-term behaviour of this three-subpopulation model is so different compared to that of the two-subpopulation model, even though the two models are based on the same birth and death rate figures. Recall that the two-subpopulation model predicted long-term population growth at a ratio of 1.003 per year (that is, at a proportionate growth rate of $0.003 = 0.3\%$ per year), while the three-subpopulation model in this activity predicts that eventually the population will decline slowly, at a ratio of about 0.999 per year (a proportionate growth rate of -0.1% per year).

A subtle point that these models do not take into account is that the age distribution *within* a subpopulation may change with time. The three-subpopulation model predicts that the number of elderly will increase as a proportion of all adults (non-juveniles). In the two-subpopulation model, the birth rate proportion is applied to the *whole* adult population, and so with time less account is taken of the fact that an increasing proportion of these adults are over the age of 65, and hence unlikely to reproduce. In the three-subpopulation model, this factor is taken into account and effectively reduces the birth rate of the population as a whole with time, by comparison with the two-subpopulation model.

In practice, models that subdivide the population into much narrower age bands (covering, say, five years each) are used. This helps to avoid such problems.

Mathcad notes

The expressions (normally seen in the middle placeholder of each vertical axis) used to plot the graphs at the bottom of pages 3 and 4 of the worksheet are not shown. Hiding the expressions for the three vertical axes enables us to display the three graphs side by side.

They were hidden by first clicking in each graph to select it, then choosing **Graph ▶ X-Y Plot...** from the **Format** menu and clicking in the check box to 'Hide arguments' on the 'Traces' tab in the 'Formatting Currently Selected X-Y Plot' option box. However, if you do this, then it is important to add a title to the graph and perhaps to add axis labels too. You can use the settings on the 'Labels' tab in the option box to add these features.

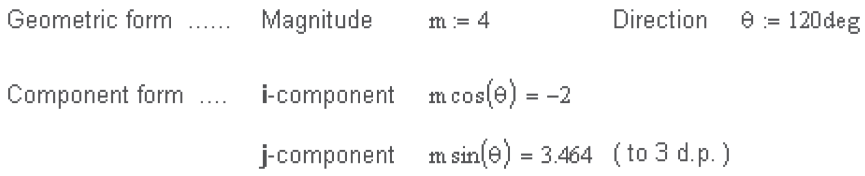
Now close Mathcad file 121B2-03.

Chapter B3

The main subject matter of this chapter is vectors, in column, component and geometric form, and their application to displacements, velocities and forces. The chapter also includes the Sine Rule and Cosine Rule, together with applications of their use.

There are no specified computer activities to accompany Chapter B3. However, there are possible applications of Mathcad in the context of the chapter, and you are encouraged to put Mathcad to use where it seems appropriate and convenient to do so. Here are some specific suggestions on where Mathcad might be applied.

◇ If a vector is given in geometric form, in terms of a magnitude and direction, then Mathcad can be used to find its components. This is illustrated below for the case where the vector has magnitude $m = 4$ and direction $\theta = 120°$.



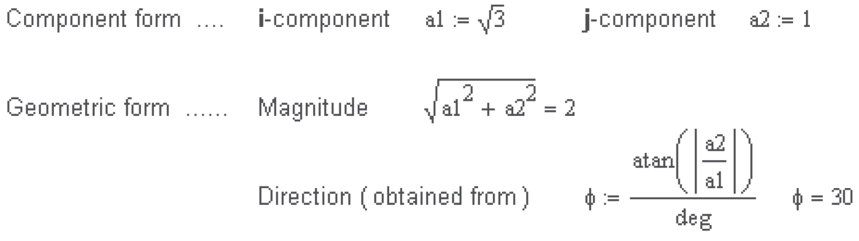
Geometric form ...... Magnitude $m := 4$ Direction $\theta := 120deg$

Component form .... **i**-component $m\cos(\theta) = -2$

j-component $m\sin(\theta) = 3.464$ (to 3 d.p.)

Note that Mathcad interprets input to any trigonometric function as being in radians. To work with angles in degrees, it is necessary to multiply by the built-in Mathcad constant $deg = \pi/180$, as shown above. (This converts from degrees to radians.)

◇ If a vector is given in component form, then Mathcad can be used to find its magnitude and to assist in finding its direction. This is illustrated below for the vector whose component form is $\sqrt{3}\,\mathbf{i} + \mathbf{j}$, with corresponding components $a1 = \sqrt{3}$ and $a2 = 1$.



Component form .... **i**-component $a1 := \sqrt{3}$ **j**-component $a2 := 1$

Geometric form ...... Magnitude $\sqrt{a1^2 + a2^2} = 2$

Direction (obtained from) $\phi := \dfrac{\operatorname{atan}\left(\left|\dfrac{a2}{a1}\right|\right)}{deg}$ $\phi = 30$

Mathcad gives an error if $a1 = 0$ is used here.

The output for the angle ϕ is found using the arctan function, denoted in Mathcad by atan. This output will be in radians, and the corresponding value in degrees is obtained on dividing by *deg*, as shown. There remains the step of finding the direction θ (in the range $-180° < \theta \leq 180°$) from ϕ (between 0 and 90°). This is achieved as explained in Figure 2.5 and surrounding text in Chapter B3, and is not a task which can be performed immediately by Mathcad.

The arctan function was introduced in Chapter A3, Subsection 4.2 (along with arcsin and arccos, referred to overleaf).

◇ Use of the Sine Rule or Cosine Rule to find an angle θ of a triangle leads to an equation of the form $\sin\theta = \ldots$ or $\cos\theta = \ldots$, respectively. Such an equation can be solved for θ using the function arcsin (asin in Mathcad) or arccos (acos in Mathcad). However, you need to bear in mind that an equation $\sin\theta = \ldots$ or $\cos\theta = \ldots$ usually has more than one solution within the range $-180° < \theta \leq 180°$. The function arccos gives the solution for which $0 \leq \theta \leq 180°$, and so always provides the correct answer for an angle within a triangle. On the other hand, arcsin gives the solution for which $-90° \leq \theta \leq 90°$ or, for an angle within a triangle, $0 < \theta \leq 90°$. You then need reasoning independent of Mathcad to decide whether it is the given output or 180° minus that output which is the angle required. Further details on this point are given in Subsection 3.1 of the chapter.

As with atan, the Mathcad functions asin and acos give output in radians, which can be converted to degrees on division by *deg*.

Recall that

$$\sin(180° - \theta) = \sin\theta.$$

Mathcad notes

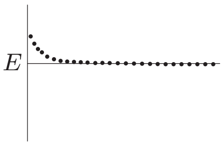
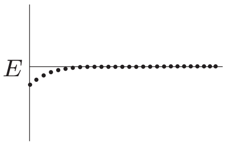
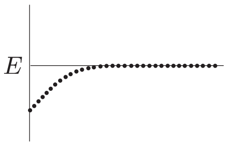
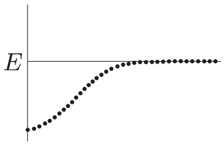
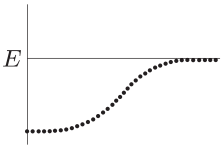
To enter Greek letters in Mathcad, you can either click on the appropriate button on the 'Greek' toolbar, or type the equivalent Roman letter followed by `[Ctrl]g`. For example, to obtain the Greek letters θ and μ, you can click on their buttons or type `q[Ctrl]g` and `m[Ctrl]g`, respectively.

Solutions to Activities

Chapter B1

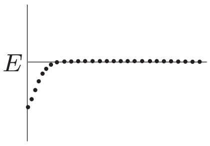
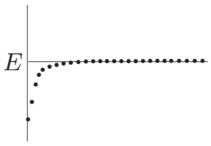
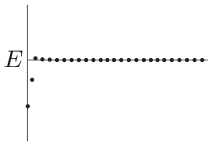
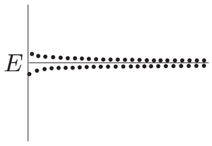
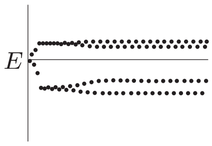
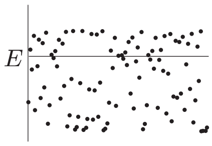
Solution 4.1

Here $r = 0.2$ and $E = 13\,300$. The descriptions, in words and sketches, are shown below. The sequences all tend to E in the long run.

P_0		
20 000	decreasing tends to E	
10 000	increasing tends to E	
5 000	increasing tends to E	
1 000	increasing tends to E slight S-shape	
100	increasing tends to E S-shape	

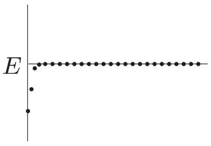
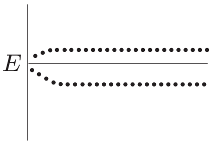
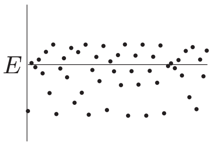
Solution 4.2

Here $P_0 = 5000$ and $E = 13\,300$. The descriptions are shown below.

r		
0.5	increasing tends to E	
1	increasing tends to E rapidly	
1.5	tends to E terms above and below E	
2	tends to E slowly terms above and below E	
2.5	has no limit takes one of four values	
3	random values between two bounds	

Solution 4.3

Again $P_0 = 5000$ and $E = 13\,300$. The descriptions are shown below.

r		
1.25	tends to E rapidly terms above and below E	
2.25	has no limit takes one of two values	
2.75	random values between two bounds	

Solution 4.4

The long-term behaviour of the sequence appears to be a 3-cycle for values of r between about 2.83 and 2.84.

Chapter B2

Solution 4.2

The answers to Task 2 are as follows:

$$\mathbf{AB} = \begin{pmatrix} 8 & 12 \\ 14 & 17 \end{pmatrix}, \qquad \mathbf{BA} = \begin{pmatrix} 26 & 6 \\ 1 & -1 \end{pmatrix},$$

$$\mathbf{A}^2 = \begin{pmatrix} 4 & 0 \\ 9 & 1 \end{pmatrix}, \qquad \mathbf{A}^3 = \begin{pmatrix} 8 & 0 \\ 21 & 1 \end{pmatrix},$$

$$\mathbf{A}+\mathbf{B} = \begin{pmatrix} 6 & 6 \\ 5 & 0 \end{pmatrix}, \qquad \mathbf{A}-\mathbf{B} = \begin{pmatrix} -2 & -6 \\ 1 & 2 \end{pmatrix},$$

$$1.5\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 4.5 & 1.5 \end{pmatrix}.$$

The following are the answers to the *optional* question. Nine products can be formed:

$$\mathbf{CC} = \begin{pmatrix} 200 & -16 \\ -64 & 8 \end{pmatrix}, \qquad \mathbf{Cx} = \begin{pmatrix} -121 \\ 26 \end{pmatrix},$$

$$\mathbf{CD} = \begin{pmatrix} -126 & 20 & 96 & -149 \\ 36 & -16 & -24 & 34 \end{pmatrix},$$

$$\mathbf{EE} = \begin{pmatrix} 68 & 67 & -117 \\ -76 & -42 & -35 \\ -107 & -11 & 158 \end{pmatrix},$$

$$\mathbf{EF} = \begin{pmatrix} 7 & 64 \\ -26 & -44 \\ -4 & 25 \end{pmatrix}, \qquad \mathbf{Ev} = \begin{pmatrix} 147 \\ -80 \\ -35 \end{pmatrix},$$

$$\mathbf{FC} = \begin{pmatrix} -8 & -8 \\ 48 & -12 \\ 2 & -7 \end{pmatrix}, \qquad \mathbf{Fx} = \begin{pmatrix} 43 \\ 7 \\ 29 \end{pmatrix},$$

$$\mathbf{FD} = \begin{pmatrix} 18 & 28 & -24 & 47 \\ -18 & 32 & 4 & 3 \\ 9 & 23 & -15 & 31 \end{pmatrix}.$$

(Note: two of the matrices involved in this exercise ($\mathbf{v}$ and $\mathbf{x}$) are vectors, and if you try to form the matrix product of each of these vectors with itself (by typing `v*v` or `x*x`), you will find that Mathcad gives a numerical answer rather than notifying of an error. While these are not valid matrix multiplications, there is another form of multiplication (known as the scalar or dot product) that can be defined between two vectors of the same size, including the case where both vectors are the same. (You will see this product introduced if you study other modules.) Mathcad implements this scalar product by employing the same symbol '$*$' that is used for other forms of multiplication, including matrix multiplication. This explains why a numerical answer can be obtained in these cases, even though matrix multiplication is not involved.)

Solution 4.3

(a) The total population and both of the subpopulations increase over the 50-year period; the number of juveniles increases by about 2 million, and the number of adults increases by about $6\frac{1}{2}$ million. The ratio of successive total populations decreases, but its graph flattens out. It appears to tend to a limit just above $1.002\,73$.

(b) Yes, the proportions of juveniles and adults both appear to tend to limits, 0.197 and 0.803, respectively (to 3 decimal places), so they are eventually fairly constant. Therefore the answer to the question is that there will be a constant proportion of juveniles joining the potential workforce, in fact, $0.197/15 \simeq 0.013$ or 1.3% of the population each year.

Changing N from 50 to 100 confirms these results.

Solution 4.4

(a) (i) Yes, the ratio of successive total populations decreases and appears to tend to a limit between 1.0027 and 1.0028.

(ii) The number of juveniles increases over the 50-year period of prediction, faster at first. The number of adults decreases at first before gradually increasing. The model predicts an increase of about 7 million juveniles and an increase of about 1 million adults by the end of the 50-year period of prediction.

(iii) The proportion of juveniles increases and appears to tend to a limit just below 0.2. The proportion of adults decreases and appears to tend to a limit just above 0.8.

(b) (i) Yes, the ratio of successive total populations increases and appears to tend to a limit just above 1.0027.

(ii) The number of juveniles decreases at first before gradually increasing. The number of adults increases, faster at first. The model predicts a decrease of about 5 million juveniles and an increase of about 15 million adults by the end of the 50-year period of prediction.

(iii) The proportion of juveniles decreases and appears to tend to a limit just below 0.2. The proportion of adults increases and appears to tend to a limit just above 0.8.

(c) For both sets of initial values for the subpopulation sizes, the ratio of successive total populations appears to tend to a limit between 1.0027 and 1.0028, and the proportions of juveniles and adults seem to tend to limits just below 0.2 and just above 0.8, respectively. The actual numbers of juveniles and adults are different for the two cases, but the long-term behaviour of the model does not appear to depend on the initial values of the subpopulation sizes.

Solution 4.5

(a) (i) The ratio of successive total populations appears to tend to a limit just below 0.999. This ratio is less than one, so the prediction is for a slight year-on-year decrease in the total population.

(ii) The proportion of juveniles increases at first, then decreases slightly; it looks as if the graph may flatten out and tend to a limit. The proportion of workers decreases, faster at first; the graph appears to flatten out and tend to a limit. The proportion of elderly increases, faster at first; this graph also appears to flatten out and tend to a limit.

(b) By the end of the 100-year period of prediction, the ratio of successive total populations still appears to tend to a limit, between 0.998 and 0.999. The proportions of the subpopulations all seem to have graphs that flatten out and tend to limits, at about 0.1925 for the juveniles, just below 0.6 for the workers, and at about 0.21 for the elderly.

(c) The model predicts that, at the end of the 50-year period, the ratio of the elderly to the workers supporting them is about 1:3. It appears therefore that a stable and sustainable situation may be reached, in terms of support of the elderly by the workers.

(For example, if a state pension is to be one half of the average working wage, then the fraction of each worker's wage needed to support pensions for the elderly is predicted to be $\frac{1}{3}$ of $\frac{1}{2}$, that is, $\frac{1}{6}$, or about 17%.)

Computer Book C

Continuous Models

Guidance notes

This computer book contains those sections of the chapters in Block C which require you to use Mathcad. Each of these chapters contains instructions as to when you should first refer to particular material in this computer book, so you are advised not to work on the activities here until you have reached the appropriate points in the chapters.

In order to use this computer book, you will need the following Mathcad files.

Chapter C1

121C1-01 Differentiation

Chapter C2

121C2-01 Integration as a limit of summations (Optional)

Chapter C3

121C3-01 Direction fields and solution curves
121C3-02 Euler's method

Instructions for installing these files onto your computer's hard disk, and for opening them, are given in Chapter A0.

The computer activities for Chapters C1 and C2 require you also to work with Mathcad worksheets which you have created yourself.

Activities based on software vary both in nature and in length. Sometimes the instructions for an activity appear only in the computer book; in other cases, instructions are given in the computer book and on screen. Feedback on an activity is sometimes provided on screen and sometimes given in the computer book.

For advice on how each computer session fits into suggested study patterns, refer to the Study guides in the chapters.

Chapter C1, Section 5
Optimisation with the computer

The Mathcad activities for Chapter C1 have only one prepared file associated with them. This file explains how to differentiate with the computer, which is the topic of Subsection 5.1. To solve optimisation problems, in Subsection 5.2, you will be given guidance on how to create your own Mathcad worksheets.

By the end of this section, you should be able to use Mathcad to carry out the process of differentiation and to solve optimisation problems, and also be a more skilled and independent Mathcad user.

5.1 Finding derivatives

In Mathcad, derivatives are found using the $\frac{d}{dx}$ operator. This operator can be used both to find a general formula for the derivative of a function f, and to find the numerical value of the derivative of f at a particular point.

For example, the derivative of the function $f(x) = \frac{1}{5}x^2$ can be described by the formula $f'(x) = \frac{2}{5}x$, which holds for all real values of x. Mathcad can obtain this expression for the derivative by using the $\frac{d}{dx}$ operator and its symbolic commands, as described in Activity 5.1. By defining a value for x before these calculations, $x = 3$ say, the value of the derivative at that particular point can be found as $f'(3) = \frac{2}{5} \times 3 = 1.2$, as you will see in Activity 5.2.

Activity 5.1 Finding a formula for the derivative

Remember to create your own working copy of the file.

Open Mathcad file **121C1-01 Differentiation**. Page 1 introduces the worksheet. Work through page 2, and then carry out Task 1 on page 3.

Solutions are given on page 84.

Comment

◇ Each part of Task 1 concerns a function that you were asked to differentiate in the main text. In the solutions, where the expression obtained using Mathcad differs in appearance from that obtained by hand, both expressions are given. If you cannot see immediately why such a pair of expressions are equivalent, it is a good idea to copy down one expression and use algebra to verify that it can be rearranged to give the other.

◇ Make sure that the variables entered in the two placeholders of the $\frac{d}{dx}$ operator match. For example, if you mistakenly try to evaluate symbolically the expression

$$\frac{d}{dt}\cos(4x),$$

then you will obtain the answer 0.

◇ Brackets are necessary when entering an expression of more than one term into the right-hand placeholder of the $\frac{d}{dx}$ operator. However, if the expression $x^3 - 6x^2 - 15x + 54$ is entered into the right-hand placeholder, then Mathcad will automatically add appropriate brackets when the first plus or minus sign is typed.

- ◇ Sometimes the expression for a derivative obtained by Mathcad can be 'improved' by simplifying it. In place of symbolic evaluation (→), either of the symbolic keywords 'simplify' and 'factor' can be applied to obtain a derivative. The outcome from these may or may not be the same as that from using →, but will sometimes be in a more convenient form. However, what Mathcad regards as 'simpler' may not necessarily seem so to a human observer!

These symbolic keywords were introduced in Chapter A0, file 121A0-05. See also *A Guide to Mathcad* for further details.

Mathcad notes

The expressions e^x and $\exp(x)$ are equivalent in Mathcad, but the former is always used in the output of symbolic calculations, irrespective of the form used for input. Similarly, the square root sign appears in output rather than the power $\frac{1}{2}$ (although $x^{3/2}$ is used rather than $x\sqrt{x}$, etc.).

Remember that Mathcad notes are *optional*.

Activity 5.1 showed how to use the $\frac{d}{dx}$ operator and symbolic evaluation to obtain an algebraic expression for the derivative. This replicates what you might do by hand, but Mathcad can also be used to differentiate functions that would be rather complicated to do by hand.

We turn next to how Mathcad can be used to find the numerical value for the derivative at a particular point.

Activity 5.2 Evaluating the derivative at a point

Turn to page 4 of the worksheet, and carry out Task 2.

You should still be working with Mathcad file 121C1-01.

Solutions are given further down page 4 of the worksheet.

Comment

- ◇ Note the comments made towards the bottom of page 4 of the worksheet. Once a value has been defined for the differentiation variable (x, say), then the $\frac{d}{dx}$ operator can be evaluated either symbolically ($\frac{d}{dx}(\ldots) \rightarrow$) or numerically ($\frac{d}{dx}(\ldots) =$) to find the numerical value of the derivative at that particular value of x.

  These two evaluation methods usually give the same numerical value, but Mathcad calculates the two results in different ways. For symbolic evaluation, Mathcad finds a general formula for the derivative and then evaluates this formula for the particular value of x, whereas for numerical evaluation, Mathcad uses a numerical algorithm to find an approximate decimal value.

Essentially, this involves a direct application of the definition of derivative, as given by equation (1.3) in Chapter C1. See also the second Mathcad note overleaf.

- ◇ If it is desired to 'turn off' a numerical value assigned previously to x (say), so as to obtain an algebraic expression from symbolic evaluation of a derivative with respect to x, this can be achieved by inserting the assignment $x := x$ just before the derivative.
- ◇ If the expression $\frac{d}{dx}f(x)$ is entered, where a definition for the function $f(x)$ is provided earlier in the worksheet, then Mathcad can find either a symbolic derivative for $f(x)$ or the numerical value of this derivative at any specified point, just as before.

Mathcad notes

◇ When you enter → to evaluate an expression symbolically, or = to evaluate it numerically, it doesn't matter where on the expression the blue editing lines are. All that matters is that the expression is complete, with every placeholder filled in. These two evaluation methods also behave in a similar way if a change is made to the worksheet, above or to the left of a calculation. In automatic calculation mode (the default), the result of a calculation involving either → or = is updated automatically, while in manual mode, you can press the `[F9]` key to update the result.

To change from automatic to manual calculation mode, or vice versa, first choose **Calculate** from the **Tools** menu, then click on **Automatic Calculation**.

◇ When evaluating a derivative numerically ($\frac{d}{dx}(\ldots) =$), you must define earlier in the worksheet the point at which the derivative is to be found; for example, $x := 3$. Mathcad then uses a numerical algorithm to obtain an approximation to the exact value of the derivative at that point, which is usually accurate to 7 or 8 significant figures. Very occasionally the method fails, in which case the derivative is highlighted in red. Clicking on this expression reveals the error message 'This calculation does not converge to a solution.'.

Without this prior definition, x appears in red in the $\frac{d}{dx}$ operator, as an undefined variable.

Now close Mathcad file 121C1-01.

Sometimes you may want to use Mathcad to find a second-order derivative. One way to do this is by using the $\frac{d}{dx}$ operator twice. You can enter the $\frac{d}{dx}$ operator in your worksheet, then enter the $\frac{d}{dx}$ operator again into the right-hand placeholder, and then fill in all the placeholders appropriately. For example, Mathcad gives the result

$$\frac{d}{dx}\left(\frac{d}{dx}(x^3 - 6x^2 - 15x + 54)\right) \;\rightarrow\; 6x - 12.$$

Another way to find a second-order derivative in Mathcad is to use the $\frac{d^n}{dx^n}$ operator, for which there is a button on the 'Calculus' toolbar. You should enter '2' in the bottom index placeholder (which will cause a 2 to appear in the top placeholder as well), and fill in the other placeholders just as for the $\frac{d}{dx}$ operator. You can evaluate the resulting expression either symbolically or numerically, in the same way as for expressions involving the $\frac{d}{dx}$ operator. For example, Mathcad gives the result

The keyboard alternative is `[Ctrl]?` (given by the three keys `[Ctrl]`, `[Shift]` and `/`).

$$\frac{d^2}{dx^2}(x^3 - 6x^2 - 15x + 54) \;\rightarrow\; 6x - 12.$$

5.2 Optimisation

Optimisation involves finding the greatest or least value taken by a function on an interval. In the main text you saw how to apply the Optimisation Procedure by hand, but all the calculations required can also be performed using Mathcad. In some cases, Mathcad can be applied to check calculations already done by hand, while in other cases, calculations can be carried out that are too complicated to do by hand.

See Chapter C1, Subsection 2.3.

In this subsection you will solve two optimisation problems, the first a minimisation problem and the second a maximisation problem.

The orienteer's problem

The orienteer's problem was described in Computer Book A. The orienteer starts at a point in a forest and needs to reach a final point on a path. The problem is to find where the orienteer should aim to join the path, where the joining point is x km from a fixed point O on the path, in order to minimise the time taken overall. For the particular data given, this involves finding the value of x that gives the least value of the journey time (in hours)

See Computer Book A, Chapter A3, Subsection 5.5.

$$f(x) = 0.125\sqrt{1+x^2} + 0.0625(2-x) \quad (x \text{ in } [0,2]).$$

The graph of $y = f(x)$ is shown in Figure 5.1.

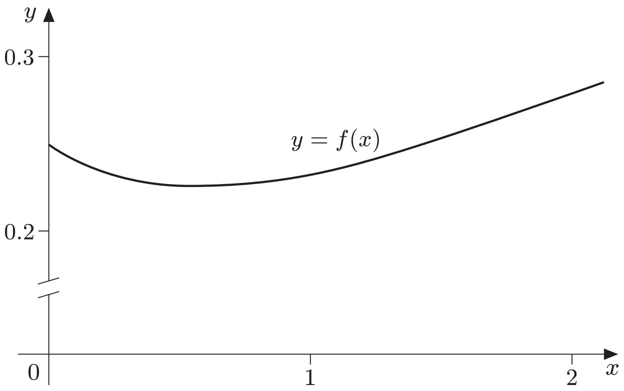


Figure 5.1 Graph of the function for the orienteer's problem

In Computer Book A, you found an approximate solution to this problem using Mathcad, by zooming in on the graph and applying the graph trace tool. You are now in a position to solve the problem more accurately, using differentiation.

In Activity 5.3 you are asked to use the Optimisation Procedure to solve the orienteer's problem. In place of a prepared Mathcad file for this activity, you are guided through creating your own Mathcad worksheet. This includes a requirement to enter some explanatory text, to make the worksheet more comprehensible.

The instructions for this activity may look lengthy, but several of them describe general Mathcad techniques that can save you time and effort, now and in the future. The instructions are given mostly in a mouse/click way, but details of keyboard alternatives are provided throughout. If you wish to use the toolbar buttons, then the activity makes use of buttons on the 'Calculator', 'Calculus', 'Boolean' and 'Symbolic' toolbars.

To open a toolbar, either click on the appropriate button on the 'Math' toolbar, or use the **View** menu, **Toolbars**.

Activity 5.3 Minimising the orienteer's journey time

A Guide to Mathcad contains detailed information on creating and editing your own worksheets. See also Activity 2.3 in Chapter A0.

The following instructions guide you through creating a Mathcad worksheet to carry out the Optimisation Procedure, in order to solve the orienteer's problem. Part (b) describes how to enter a title in the worksheet. You should similarly enter any other text that you think is appropriate. For example, you could include a line of text before each step of the Optimisation Procedure, to explain what you are doing.

If you have just started Mathcad running, then there is no need to do this, as it automatically starts with a new (Normal) worksheet.

(a) Begin by creating a new worksheet, as follows. Select the **File** menu and choose **New...** . In the list of templates that appears, **Normal** should be selected by default. If not, click on it. Then click on the **OK** button to create a new (Normal) worksheet. (Alternatively, type `[Ctrl]n`, or click on the 'New' button on the standard toolbar.)

Any text that you enter can later be moved, or deleted, as you wish.

(b) Enter a title at the top of your worksheet. To do this, click to position the red cross cursor in an appropriate place, and choose **Text Region** from the **Insert** menu. (Alternatively, type a double-quote `"`, given by `[Shift]2`.) Then type a suitable title, for example, `Optimisation - the orienteer's problem`, in the text box. To finish, click anywhere outside the text box or press `[Ctrl][Shift][Enter]`. If you need to edit the text later, simply click on it.

If you have difficulty in following these instructions, then you may like to look ahead to the first item in the Comment on the next page, which shows what should eventually appear on the screen.

(c) Enter a definition of the function to be optimised, which is

$$f(x) = 0.125\sqrt{1+x^2} + 0.0625(2-x).$$

You can use the keyboard and buttons on the 'Calculator' toolbar to enter this, or just type the key sequence

`f(x):0.125*\1+x^2[Space][Space][Space][Space]+0.0625*(2-x)`

The `\` (backslash) keystroke creates the square root sign.

(d) Now you are ready to solve the orienteer's problem, by carrying out the three steps of the Optimisation Procedure, as follows.

Step 1: Find the stationary points of f

Note that this button gives the special equals sign, which equates the expressions on either side of it. (It has thicker lines than the equals sign obtained from the 'Evaluate Numerically' = button on the 'Calculator' toolbar.)

Using the $\frac{d}{dx}$ button on the 'Calculus' toolbar, enter the expression $\frac{d}{dx}f(x)$ and select the whole expression. Then click on the 'Equal to' = button on the 'Boolean' toolbar, and enter `0` (zero). (Alternatively, just type `?x[Tab]f(x)[Space][Ctrl]=0`. The `?` (question mark, given by `[Shift]/`) enters the $\frac{d}{dx}$ operator, `[Tab]` moves between the placeholders, `[Space]` expands the selection and `[Ctrl]=` gives the special equals sign.)

The stationary points are the points where the derivative of the function f is zero, that is, the solutions of the equation just typed in. Hence (with the vertical blue editing line still at the right-hand end of this equation), click on the 'solve' button on the 'Symbolic' toolbar. (Alternatively, just type `[Ctrl][Shift].solve`.) Finally, click elsewhere on the page, or press `[Enter]`.

You should see the solution 0.577 350 ... appear to the right of '→'. This is the single value of x at which the function f has a stationary point.

Step 2: Evaluate f at each of the relevant points

Use Mathcad to evaluate the original expression $f(x)$ at the two endpoints of the interval, $x = 0$ and $x = 2$, and at the stationary point between them. For example, type `f(0)=` to evaluate $f(0)$. It is sufficient to input the value for the stationary point to three decimal places, $x = 0.577$, or you can use copy and paste to input all 20 digits given in the solution if you wish!

A Comment item below describes in detail how to use the copy and paste facilities.

Step 3: Choose the optimum value

Identify the least of the three function values that you have found. This is the minimum value of $f(x)$ within the given interval.

Remember that the orienteer's problem is a minimisation problem.

You could finish at this point, but it is a good idea to record your conclusions in the worksheet. For example, you could enter the text

```
The least value of f(x) on [0,2] is ?  (at x=?).
```

replacing the question marks with the values that you found. You can add more text if you wish, but the Optimisation Procedure for the orienteer's problem is now completed.

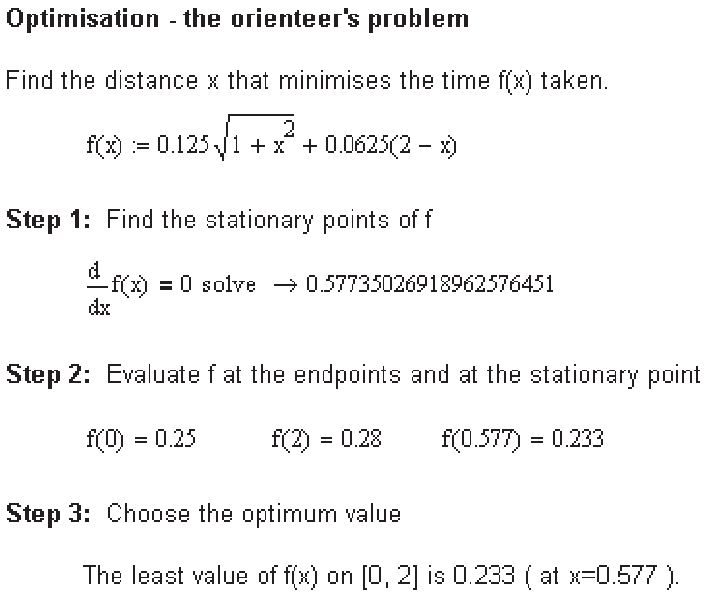
(e) Finally, save your file, by choosing **Save As...** from the **File** menu. You will need to give it a suitable name that indicates the contents, for example, *my121C1-orienteer*. This will help you to organise and identify your files.

A solution to part (d) is given on page 84.

Comment

◇ The Mathcad worksheet should now look something like this:

Your worksheet should contain the calculations shown, but may have different text.



Optimisation - the orienteer's problem

Find the distance x that minimises the time f(x) taken.

$f(x) := 0.125\sqrt{1 + x^2} + 0.0625(2 - x)$

Step 1: Find the stationary points of f

$\frac{d}{dx}f(x) = 0 \text{ solve } \rightarrow 0.57735026918962576451$

Step 2: Evaluate f at the endpoints and at the stationary point

$f(0) = 0.25$ $f(2) = 0.28$ $f(0.577) = 0.233$

Step 3: Choose the optimum value

The least value of f(x) on [0, 2] is 0.233 (at x=0.577).

- ◇ It would have been possible to solve the equation $\frac{d}{dx}f(x) = 0$ as described above but without explicitly entering '`= 0`'. When an expression does not contain an equals sign, the symbolic keyword 'solve' will find values of the selected variable for which the expression is equal to zero. (However, it is clearer to read such a line of the worksheet with '`= 0`' included.)
- ◇ If you want to evaluate $f(0.577\,350\ldots)$ using all 20 decimal places given in the solution for the stationary point, then rather than enter the number from scratch, you can take advantage of Mathcad's copy and paste facilities. To do this, click anywhere in the solution (the horizontal blue editing line extends to select the entire number, regardless of where the vertical editing line is positioned within it) and choose **Copy** from the **Edit** menu. Then type `f()` , and paste this number into the empty placeholder between the brackets by choosing **Paste** from the **Edit** menu. Lastly, click or type `=` to evaluate the function for this value. (The keyboard alternatives for copy and paste are `[Ctrl]c` and `[Ctrl]v`, respectively.)

Copy and paste is a handy way of avoiding the need to retype awkward or lengthy expressions.

Mathcad notes

When working symbolically, a decimal point in the input expression triggers a decimal result, with up to 20 decimal places! However, when working numerically (evaluating an expression using `=`), the number of decimal places displayed depends on the setting for 'Number of decimal places' (**Format** menu, **Result...**, 'Number Format' tab). By default, the results of numerical calculations are displayed to three decimal places.

Now close the Mathcad file that you have created and saved.

A traffic planning model

Traffic planners wish to set up a mathematical model to describe how the volume flow rate of traffic (that is, the number of vehicles which pass a fixed point in a given time) varies with the average velocity of vehicles along a single lane of a road. The eventual purpose of this model is to advise on how traffic flow can be maximised.

The planners assume that each vehicle moves with a constant velocity $v\ \mathrm{m\,s^{-1}}$. On the basis of a subsidiary model and many observations, they estimate that, on average, each driver maintains a distance $v + 0.02v^2$ metres between the front of their vehicle and the back of the vehicle immediately ahead. The average length of a vehicle is estimated to be 5 metres. The model therefore represents the situation as shown in Figure 5.2.

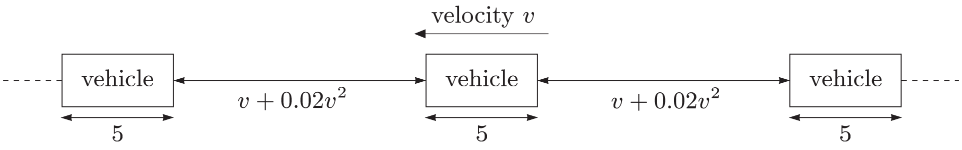


Figure 5.2 Representation of traffic flow

The distance between the fronts of successive vehicles is $5 + v + 0.02v^2$ metres. Each vehicle, travelling at velocity v m s^{-1}, covers this distance in $(5 + v + 0.02v^2)/v$ seconds. It follows that the volume flow rate $f(v)$ of traffic (in vehicles per second past a fixed point) is given by

$$f(v) = \frac{v}{5 + v + 0.02v^2}.$$

This formula is to apply for $0 \leq v \leq 35$.

A velocity of 35 m s^{-1} is about 126 kilometres per hour or 78 miles per hour.

The traffic planners seek the maximum value of $f(v)$ for v in the interval $[0, 35]$. If such a maximum can be found, it can be used to provide an advised speed of travel on the road.

In the next activity you are asked to create a Mathcad worksheet once more, to find both the greatest value of the volume flow rate function $f(v)$ and the value of v for which this occurs.

Activity 5.4 Maximising the volume flow rate of traffic

(a) Create a new (Normal) worksheet, as in Activity 5.3(a). (Alternatively, you may prefer to work with the file that you created for the orienteer's problem in Activity 5.3. To do this, open the file, make a working copy of it using a *different* file name, for example, *my121C1-traffic*, then edit the text and expressions in the worksheet.)

(b) Enter a title at the top of your worksheet, for example, `Maximising the volume flow rate of traffic.`

(c) Solve the volume flow rate problem by finding the maximum value of the function

$$f(v) = \frac{v}{5 + v + 0.02v^2}$$

on the interval $[0, 35]$ and the corresponding value of v. Do this by carrying out the three steps of the Optimisation Procedure, just as you did in Activity 5.3(d).

Remember that this is a maximisation problem.

(d) State your answer in a form which is appropriate in the context of the traffic planning model.

Solutions to parts (c) and (d) are given on page 84.

Comment

Having defined the function $f(v)$ for the volume flow rate of traffic in your worksheet, you may wish to display a graph of it. You can do so by defining a suitable graph range, for example, $v := 0, 0.1 .. 35$, and then plotting $f(v)$ against v. While the graph confirms the maximum value at $v \simeq 16$, it also shows that there is little difference in the value of $f(v)$ between $v = 10$ and $v = 20$.

Save your file, for example as my121C1-traffic. Then close this file.

Chapter C2, Section 5
Integration with the computer

There is only one prepared Mathcad file for this section, and that comes towards the end and is optional. As in the later computer activities for Chapter C1, you will, for the most part, be creating your own worksheets here and using Mathcad directly. Subsection 5.1 shows how to find indefinite integrals in Mathcad, while Subsection 5.2 covers definite integrals.

5.1 Finding indefinite integrals

In Mathcad, indefinite integrals are found using the $\int$ operator. Like the $\frac{d}{dx}$ operator, the $\int$ operator can be used with Mathcad's symbolic commands, to find an algebraic expression for an integral of a given function.

In this subsection you are invited to find indefinite integrals for a variety of functions. The first activity provides an introduction to finding indefinite integrals using Mathcad.

Activity 5.1 How to find indefinite integrals

In this activity you will use Mathcad to find the indefinite integral of x^2.

The buttons referred to below are on the 'Calculus' and 'Symbolic' toolbars. If you wish to use these and they are not already visible, then either click on the appropriate buttons on the 'Math' toolbar, or select the **View** menu, **Toolbars** and choose **Calculus**, then repeat and choose **Symbolic**.

If necessary, see Chapter C1, Activity 5.3(a) on page 64 of this computer book.

Be careful not to confuse the $\int$ button with the $\int_a^b$ button, which is used for finding definite integrals (as you will see later).

(a) Create a new (Normal) worksheet.

(b) Enter the $\int$ operator in your worksheet, either by clicking on the $\int$ button on the 'Calculus' toolbar, or by using the keyboard alternative `[Ctrl]i`.

(c) Enter the expression to be integrated, x^2, in the first placeholder after the integral sign. (This expression is called the integrand.) Then enter the variable of integration, x, in the placeholder after the 'd'.

(d) Click on the → button ('Symbolic Evaluation') on the 'Symbolic' toolbar, or use `[Ctrl].`, the keyboard alternative. Then click elsewhere on the page, or press `[Enter]`, to obtain the integral. Check that the answer provided by Mathcad is what you expect.

(e) Now go through the same procedure to evaluate the integral $\int u^2\,du$.

(f) If you wish to save your work, then select the **File** menu and use **Save As...** to name and save your worksheet. (It is a good idea to insert a title in your worksheet. If you need to create space for this, then do so by positioning the red cross cursor at the top of the worksheet and pressing `[Enter]` to insert as many blank lines as required.)

Comment

- Notice that the $\int$ operator in Mathcad gives only *an* integral of the integrand. It does not give *the indefinite* integral because it does not add an arbitrary constant.
- The outcomes from integrating $\int x^2\,dx$ and $\int u^2\,du$ demonstrate that the form of the indefinite integral depends on the nature of the function being integrated but not on the choice of symbol for the variable of integration.

Mathcad notes

- In part (e), it is sufficient simply to edit the integral expression $\int x^2\,dx$, replacing each 'x' by a 'u'.
- You can also find an integral of a function f by first defining f and then evaluating symbolically the expression $\int f(x)\,dx$.
- The behaviour of the $\int$ operator differs somewhat from that of the $\frac{d}{dx}$ operator, used in Chapter C1. The $\int$ operator *cannot* be evaluated numerically (by typing =), since this makes no sense in the context of finding an indefinite integral. If you try this, after setting a value for the variable of integration, then the integral is highlighted in red. Clicking on the integral reveals the error message 'This operator must be evaluated symbolically.'.

  However, once x has been assigned a value, symbolic evaluation of an indefinite integral with respect to x does give a numerical answer in Mathcad, just as for a derivative. This unwanted behaviour can be prevented by typing $x := x$ just before the indefinite integral.

A similar usage of $x := x$ was referred to in the context of derivatives, on page 61.

In each of the remaining activities in this subsection, you can either continue using the worksheet from the previous activity, or close that file and create a new worksheet. You will need to decide whether you wish to save your work.

In the next activity, you are asked to use Mathcad to find the indefinite integrals of two functions that you integrated by hand in the main text, and of a third function which you could also integrate by hand.

Activity 5.2 Finding indefinite integrals

Use Mathcad to find each of the following indefinite integrals.

(a) $\displaystyle\int\left(\frac{1}{x}+e^{3x}\right)dx \quad (x>0)$ (b) $\displaystyle\int\left(\frac{3}{y^4}+5\sin(5y)\right)dy \quad (y>0)$

(c) $\displaystyle\int(a+\cos(ax))\,dx$ (where a is a non-zero constant)

Solutions are given on page 84.

If necessary, refer to the instructions in Activity 5.1(b)–(d).

If you arrive at an incorrect answer to part (c), then see the Mathcad notes below.

Comment

- After the '$+\,c$' has been added, the Mathcad answers to parts (a) and (b) agree with those obtained in the main text, and that for part (c) agrees with the answer found by applying Table 1.1.
- Mathcad gives answers here that agree with those obtained by hand, even when the additional constraints on the variables that accompany these integrals (for example, the condition $x > 0$ in part (a)) are not entered. For this reason, it is necessary to bear in mind that the symbolic manipulations performed by Mathcad might not be valid in all circumstances, and that the output should be interpreted with care.

Additional constraints can be specified in Mathcad, using the symbolic keyword 'assume'. See Mathcad Help for details.

Mathcad notes

It is safest to enter all products explicitly in Mathcad, by clicking on the 'Multiplication' × button on the 'Calculator' toolbar or by typing `*`. If you enter `3x` in part (a), or `5sin` or `5y` in part (b), then Mathcad will assume that you intended to enter a product, and will insert the multiplication for you. However, Mathcad will *not* help in this way if you enter `ax` rather than `a*x` in part (c). In this case, Mathcad assumes that you have entered a single variable name, 'ax', rather than the product of the variables a and x.

The next activity contains four integrals which you found by hand in Section 2, followed by three further integrals (in parts (e)–(g)) that cannot be found by hand simply on the basis of what is given in the chapter.

Activity 5.3 Further indefinite integrals

Use Mathcad to find each of the following indefinite integrals.

Note that $\sin^2 x$ should be input as $\sin(x)^2$; for example, type `sin(x)^2`.

(a) $\displaystyle\int (x-3)(x-1)\,dx$ (b) $\displaystyle\int \frac{2x-3}{\sqrt{x}}\,dx$

(c) $\displaystyle\int \sin^2 x\,dx$ (d) $\displaystyle\int \frac{x}{x^2+1}\,dx$

(e) $\displaystyle\int u\,e^{3u}\,du$ (f) $\displaystyle\int x^2\ln(5x)\,dx$ (g) $\displaystyle\int \frac{1}{\sqrt{9-t^2}}\,dt$

Solutions are given on page 85.

Comment

- ◇ The answers obtained from Mathcad in parts (a)–(d) are equivalent to the answers found by hand. However, they are not always given in an identical form, and in parts (a) and (b) some algebra is required to show the equivalence of the two forms.
- ◇ It is possible to use the $\frac{d}{dx}$ operator symbolically to differentiate each of the integrals obtained in this activity. This leads to an expression equivalent to the original integrand in each case, as expected. However, the output obtained is not always identical in form to that of the original integrand, and again some algebra is sometimes required to show the equivalence of the two forms.
- ◇ Recall (from the context of obtaining derivatives symbolically) that the symbolic keywords 'simplify' and 'factor' can be used as alternatives to →. When applied to integrals, either of these may again provide an outcome in a more convenient form.

See the final Comment item for Chapter C1, Activity 5.1 on page 61 of this computer book.

Parts (e)–(g) of Activity 5.3 show that Mathcad can extend your 'integration reach' beyond the types of functions that you have so far learnt how to integrate by hand. However, Mathcad cannot integrate every function. For example, if you ask Mathcad to evaluate symbolically either of the indefinite integrals

$$\int e^{-t^3}\,dt \qquad \text{and} \qquad \int \sqrt{\sin x}\,dx,$$

No expression in terms of standard functions is known for either of these integrals.

the response is for Mathcad to repeat the given integral without alteration. This is how Mathcad responds when it cannot find an integral.

5.2 Definite integrals, areas and summations

You will need Mathcad file 121C2-01 for (optional) Activity 5.8, later in this subsection. First, however, you are invited to create your own Mathcad worksheets as in the previous subsection, but now to find definite rather than indefinite integrals.

In Mathcad, definite integrals are found using the $\int_a^b$ operator. This operator can be used either symbolically or numerically. If the operator is used *symbolically*, then Mathcad finds an algebraic expression for an integral, evaluates this expression at the upper and lower limits of integration, and subtracts the second value from the first to find the answer. This is the same as the usual approach to finding definite integrals by hand. If the operator is used *numerically*, then Mathcad does not find an algebraic expression, but instead uses a numerical algorithm to find an approximate value for the definite integral.

In each of Activities 5.4–5.7 below, as in Subsection 5.1, you can either continue using the worksheet from the activity before, or close that file and create a new worksheet. You will need to decide whether you wish to save your work.

Activity 5.4 introduces the symbolic use of the $\int_a^b$ operator.

Activity 5.4 How to evaluate definite integrals

In this activity you will use Mathcad to evaluate the definite integral

$$\int_2^3 \frac{1}{x}\,dx.$$

(a) Enter the $\int_a^b$ operator in your worksheet, either by clicking on the $\int_a^b$ button on the 'Calculus' toolbar, or by using the keyboard alternative `&` (the ampersand sign, given on the keyboard by `[Shift]7`).

(b) Enter the integrand, the variable of integration, and the upper and lower limits of integration in the appropriate placeholders.

The `[Tab]` key provides a good way of moving around the placeholders.

(c) Click on the → button ('Symbolic Evaluation') on the 'Symbolic' toolbar, or use `[Ctrl].`, the keyboard alternative. Then click elsewhere on the page, or press `[Enter]`. You should obtain the answer

$$\ln(3) - \ln(2).$$

(d) Click anywhere on this answer, and then enter `=`, either by clicking on the `=` button on the 'Calculator' toolbar or by typing `=`, to evaluate it numerically. You should obtain the answer 0.405, which is the value of $\ln 3 - \ln 2$ to three decimal places.

Comment

Evaluating symbolically an expression involving the $\int_a^b$ operator gives an expression which is an *exact* answer (unless the original expression contains a decimal point; see the second Mathcad note below). In the example in this activity the expression is $\ln(3) - \ln(2)$. You can display the decimal value of such an expression by evaluating it numerically. This is done by selecting the expression, and then entering `=`.

Mathcad notes

◇ When you evaluate numerically an expression in Mathcad, the number of decimal places displayed is determined by the value of 'Number of decimal places'. The default value of this is 3, but you can change it by choosing **Result...** from the **Format** menu and then the 'Number Format' tab.

◇ If a Mathcad expression involving the $\int_a^b$ operator has a decimal point in any constant in the integrand, or in either limit of integration, then evaluating the expression symbolically gives a decimal answer with up to 20 decimal places. (Such an answer is unaffected by the value of 'Number of decimal places'.) For example, evaluating symbolically the integral below in Mathcad gives the outcome

$$\int_2^3 \frac{1.0}{x}\,dx \;\rightarrow\; 0.40546510810816438198.$$

The next activity provides further practice in using Mathcad to evaluate definite integrals symbolically.

Activity 5.5 Evaluating definite integrals

If necessary, refer to the instructions in Activity 5.4.

Each of parts (a)–(d) below gives a definite integral that you were asked to evaluate by hand in the main text. In each case, use Mathcad to evaluate the definite integral symbolically to obtain an exact answer, and then evaluate this answer numerically, to display it as a decimal value.

Remember that π can be obtained from the 'Calculator' or 'Greek' toolbar, or by typing `[Ctrl][Shift]p`.

(a) $\displaystyle\int_0^2 e^t\,dt$ (b) $\displaystyle\int_0^{\pi/4} (\cos(5x) + 2\sin(5x))\,dx$

(c) $\displaystyle\int_1^2 \frac{6}{u^2}\,du$ (d) $\displaystyle\int_0^{\pi} e^t \sin t\,dt$

Solutions are given on page 85.

In the next activity you will see an example of a definite integral that Mathcad is unable to evaluate symbolically. When this happens it is often worth attempting to evaluate the integral numerically, and the activity shows you how to do this.

Activity 5.6 An awkward definite integral

(a) Use Mathcad to try to evaluate symbolically the definite integral

$$\int_0^1 e^{-t^3}\,dt.$$

At the end of Subsection 5.1 it was pointed out that Mathcad cannot evaluate the indefinite integral

$$\int e^{-t^3}\,dt.$$

You should find that the definite integral is repeated without alteration. This means that Mathcad has been unable to calculate an algebraic expression for an integral of the integrand, and so it cannot evaluate symbolically the given definite integral.

(b) Evaluate the definite integral numerically, as follows. Click anywere on the expression just created, and then enter `=`.

You should find that the answer 0.808 is displayed.

Comment

- ◇ To evaluate any definite integral numerically, you should enter it in the same way as for symbolic evaluation, then select it and enter =.
- ◇ The reason why some definite integrals can be evaluated numerically but not symbolically in Mathcad is that symbolic evaluation requires Mathcad to find an algebraic expression for an integral, whereas numerical evaluation involves the use of a numerical algorithm. The answer obtained from this algorithm is an approximation, though usually an accurate one.

Activity 5.8 indicates a possible basis for such an algorithm.

Mathcad notes

On rare occasions, the numerical method used by Mathcad for evaluating definite integrals fails to produce a value. In such a case, the integral is highlighted in red. Clicking on the integral reveals the error message 'This calculation does not converge to a solution.'.

The next activity gives you further practice in using the $\int_a^b$ operator numerically.

Activity 5.7 Evaluating definite integrals numerically

(a) Evaluate numerically the definite integral $\displaystyle\int_0^1 t^2\,dt$, as follows.

(i) Either click on the $\int_a^b$ button on the 'Calculus' toolbar, or type & to insert the definite integral operator and its placeholders.

(ii) Fill in the four placeholders appropriately.

(iii) The blue editing lines should be on the expression; if not, select some part of the integral. Then evaluate the expression numerically, by entering =.

The first two steps here are identical to those for symbolic integration. These steps would also have been required for the definite integral in Activity 5.6(b), had the expression not previously been entered.

(b) Evaluate numerically the definite integral $\displaystyle\int_{-1}^1 \sqrt{1-x^2}\,dx$.

Solutions are given on page 85.

Comment

These two definite integrals can be evaluated either numerically or symbolically. As you can check, symbolic evaluation gives $\frac{1}{3}$ for (a) and $\frac{1}{2}\pi$ for (b).

You have seen that in Mathcad many definite integrals can be evaluated either symbolically or numerically. You might wonder which it is more appropriate to invoke in any given situation. If you want an exact answer (so you can see where constants such as π feature in it, for example), or if a general result is required, such as a formula for

$$\int_0^1 \cos(kx)\,dx \quad \text{(where } k \text{ is a non-zero constant)},$$

then use the symbolic approach. If you simply want a number that is an accurate value for the definite integral, then numerical evaluation should suffice. If you want both an exact value and a decimal value for the answer, then you can first evaluate the definite integral symbolically and afterwards enter =.

For example, the values of all the definite integrals in Activity 5.5 can be found satisfactorily using numerical integration.

Integration as a limit of summations

The remainder of this subsection will not be assessed.

In Subsection 4.2, you saw that a definite integral could be approximated as closely as required by a finite sum. This was demonstrated in particular for the case in which the value of the definite integral gives the area beneath the graph of a function, and each finite sum represents the overall area of a set of rectangles. Each rectangle is based upon a subinterval, and the approximation to the definite integral improves as the number of subintervals is increased.

We assume (as in the main text) that $f(x)$ takes only non-negative values in the interval $[a, b]$.

For the area beneath the graph of the function $f(x)$ between $x = a$ and $x = b$, which is given exactly by the definite integral

$$\int_a^b f(x)\,dx,$$

the approximation based on N subintervals is

$$\sum_{i=0}^{N-1} hf(a+ih) = h\sum_{i=0}^{N-1} f(a+ih), \quad \text{where } h = \frac{b-a}{N}.$$

See equation (4.1) in Chapter C2, Subsection 4.2.

The remaining (optional) activity in this section asks you to use the prepared Mathcad file 121C2-01 to explore the relationship between these approximations to the area under the curve and the definite integral itself.

Activity 5.8 Integration as a limit of summations (Optional)

(a) Open Mathcad file **121C2-01 Integration as a limit of summations**, and read through the worksheet, which consists of a single page. The definite integral being approximated here is

$$\int_0^{40} 15\sqrt{\sin\left(\frac{\pi x}{40}\right)}\,dx,$$

whose value was sought by a process of successive approximation in Subsection 4.2.

The value of this integral is the area on the graph beneath the red curve. The value calculated for A is an estimate for this area, using approximation by rectangles based on N subintervals. The value of N is initially set to 4.

Investigate the effect of increasing the number of subintervals, N. Use in turn the following values for N:

20, 50, 100, 500, 1000, 5000, 10 000.

Compare the corresponding values obtained for A with those in the right-hand column of the table below.

This is Table 4.1 in Chapter C2, Subsection 4.2.

Number of subintervals	Sum of areas of rectangles
4	402.27
20	452.71
50	456.41
100	457.21
500	457.62
1 000	457.64
5 000	457.65
10 000	457.66

(b) You may like to use the worksheet for other definite integrals, to investigate the behaviour of area estimates A as N is increased. For example, you could investigate these estimates for the definite integral

$$\int_0^1 e^{-x^3}\,dx,$$

by first editing the worksheet so that $f(x) = e^{-x^3}$ and $b = 1$.

If you have time, look also at the behaviour of the corresponding area estimates for the definite integral

$$\int_0^{\pi/4} \tan x\,dx.$$

Comment

◇ You should find that the numerical values obtained for A in part (a) match to two decimal places the values given in the right-hand column of the table. These area estimates appear to tend to the limit 457.66 (to two decimal places), which is also the numerical value provided by Mathcad for the definite integral.

◇ The values of the definite integrals suggested in part (b) are

$$\int_0^1 e^{-x^3}\,dx = 0.808 \quad\text{and}\quad \int_0^{\pi/4} \tan x\,dx = 0.347,$$

each to three decimal places. In each case, the estimates A converge to this value as N increases.

In the first example, the area estimates are greater than the value given by the definite integral for the area under the curve, while in the second example they are smaller. You can see this illustrated on the graph by setting small values of N ($N \leq 20$, say). In the first example, the tops of the rectangles all lie above the curve, while in the second example they all lie below.

Mathcad notes

◇ The summation sign is obtained from the $\sum\limits_{n=1}^{m}$ button on the 'Calculus' toolbar, or by typing `[Ctrl]$` (for which you have to press the three keys `[Ctrl]`, `[Shift]` and `4` together).

◇ Note that the definite integral is set up in this file with integrand $f(x)$, where the function $f(x)$ is defined earlier in the worksheet. This approach is possible for either numerical or symbolic evaluation of a definite integral.

◇ The rectangles are filled in by drawing zigzag lines very close together. On the screen these lines give the appearance of a solid block of colour, but the tiny gaps between the lines may become apparent if printed.

Now close Mathcad file 121C2-01.

Chapter C3, Section 4
Differential equations with the computer

In this section you will use the computer to see how the information contained in a differential equation can be displayed graphically, and how differential equations can be 'solved' numerically and graphically, even where no formula for the solution can be found.

There are two prepared Mathcad files associated with this section. The first draws direction fields and plots solution curves (graphs of solutions) for first-order differential equations, and the second illustrates how a numerical method for obtaining the solution to an initial-value problem works in practice.

This numerical method can be applied to any differential equation of the form

As special cases, f may be a function of x alone or of y alone.

$$\frac{dy}{dx} = f(x, y),$$

where f is a known function of two variables. This form includes the types of differential equation considered in the main text, but also others.

Mathcad does not contain any symbolic facilities for solving differential equations directly, to find a formula for the solution. However, as you saw in Chapter C2, Mathcad can be used to find integrals, and finding integrals is the main constituent of the two methods for solving first-order differential equations that are introduced in the main text (direct integration and separation of variables). Hence Mathcad can be used indirectly to help solve some first-order differential equations. Mathcad can also be applied to differentiate a solution that you have already found, to check that it is indeed a solution to the given differential equation.

Differentiation was the subject of Chapter C1.

In this section, however, Mathcad will be used numerically and graphically rather than symbolically.

4.1 Direction fields and solution curves

The differential equation

$$\frac{dy}{dx} = x + y$$

cannot be solved, using the methods of this chapter, to give an equation relating x and y. However, the direction field of this differential equation provides enough information to indicate the different types of solution curve that occur. In Activity 4.1 you will see this direction field drawn by Mathcad, and will be able to ask Mathcad to plot the solution curve through any point of your choice.

Activity 4.1 The differential equation dy/dx = x + y

Open Mathcad file **121C3-01 Direction fields and solution curves**. The worksheet opens with the direction field of the differential equation

$$\frac{dy}{dx} = x + y$$

drawn for a grid of points with integer coordinates in the (x, y)-plane, for values of x between -5 and 5, and values of y between -2 and 4. For each such point (x, y) a line segment is plotted through the point, and the slope of this line segment is $f(x, y) = x + y$.

(a) Set S to 1, so that a solution curve is plotted through the point $(0, 0)$. Briefly describe the curve obtained, or make a small sketch of it. Now change the value of y_0, to obtain a solution curve through each of the following points in turn:

The initial condition is $y(x_0) = y_0$.

$$(0, 1), \quad (0, 2), \quad (0, -1), \quad (0, -2).$$

In each case, note down a brief description of the curve or make a small sketch of it.

(b) Can you group the solution curves seen in part (a) into distinct types of behaviour? How many different types of behaviour are there?

For each of the following points, try to predict from looking at the direction field which type of behaviour the solution curve through the point will exhibit:

$$(-3, -1), \quad (-1, 0), \quad (4, 2).$$

Then, by making a suitable choice of values for x_0 and y_0, use Mathcad to plot the corresponding solution curve and to confirm your prediction.

Solutions are given on page 85.

Comment

The line segments of the direction field give a good indication of where the solution curves lie.

Mathcad notes

◇ A definition for a function of two variables is created in the same way as that for one variable. For example, to define the function $f(x, y) = x + y$ in Mathcad, you could type `f(x,y):x+y`.

◇ The calculations used to draw the direction field and solution curve are 'hidden' off the page of the Mathcad worksheet to the right. The area beyond the right-hand margin of a Mathcad page (which is marked by a solid vertical line) can be used just like the rest of the worksheet. It is divided into further pages, where you can place mathematical expressions, text, graphs and pictures. You do not need to look at the calculations in this worksheet, but in general, you can view what is in the 'hidden' area by using the horizontal scroll bar to move to the right.

In the next activity you are asked to use Mathcad to plot the direction fields for two other first-order differential equations. In each case, you can try to visualise from the direction field how the solution curves behave, before plotting some of these curves.

Activity 4.2 Direction fields and solution curves

You should still be working with Mathcad file 121C3-01.

(a) Investigate the direction field and solution curves for the first-order differential equation

$$\frac{dy}{dx} = e^{\cos x} - 1,$$

as follows.

(i) First set the variable S to zero, so that no solution curve is plotted. Then enter the right-hand side for this differential equation into the definition for $f(x, y)$.

(ii) Try to predict from the direction field where the solution curves lie. What types of behaviour will the solution curves exhibit?

(iii) Set the variable S to 1 to display a solution curve, and then try different values for x_0 and y_0 to confirm the predicted behaviour of solution curves. Use your observations to describe the behaviour of the solution curves for the differential equation.

(b) Now follow the same procedure as in part (a) to investigate the direction field and solution curves for the first-order differential equation

$$\frac{dy}{dx} = xy + 1.$$

However, in this case, start by altering the scope of the grid in the y-direction, by setting $Y1 := -5$, $Y2 := 5$ and $q := 10$.

Solutions are given on page 86.

Comment

◇ The differential equation in part (a) is of the form $dy/dx = f(x)$, and so can be solved in principle by direct integration. However, it turns out not to be possible to find an algebraic expression for

$$\int (e^{\cos x} - 1)\, dx.$$

The presence of the $\cos x$ means that the slopes of the direction field repeat at horizontal intervals of 2π. (The slopes are also invariant in the vertical direction, for any given choice of x, as noted in the solution.) The alternate positive and negative slopes indicate that solution curves will undulate. The overall increasing trend is not so obvious from the direction field, but might be expected because the magnitude (steepness) of the slopes appears to be greater where the slopes are positive than where they are negative.

◇ The differential equation in part (b) cannot be solved by the methods of MST121. Using Mathcad to plot the direction field provides a good approach to visualising the types of solution curve for such a differential equation.

Now close Mathcad file 121C3-01.

4.2 Euler's method

In Subsection 4.1 you saw how a direction field can be used to visualise the information provided by a first-order differential equation. You also saw solution curves plotted on top of the direction field. It is straightforward to plot such curves for a first-order differential equation whose solutions can be expressed in terms of a simple algebraic formula. However, even where no such formula can be found, approximate solution curves for a differential equation can still be plotted. These graphs are based on a sequence of numerical estimates for solution values, which constitute a numerical solution to the differential equation. In this subsection you will see how such a numerical solution can be obtained.

This was the case for each of the differential equations in Activity 4.2.

The procedure to be described below, for obtaining an approximate numerical solution to a first-order differential equation, is known as *Euler's method*. To see how the solution is built up step by step, it is illuminating to consider the corresponding graphical construction. This involves forming a connected chain of line segments, each of which has a gradient given by the slope of the direction field at the left-hand end of the line segment, as shown in Figure 4.1.

Recall that a direction field provides a slope value at every point within a given region, and not just at the particular grid points displayed on a graph.

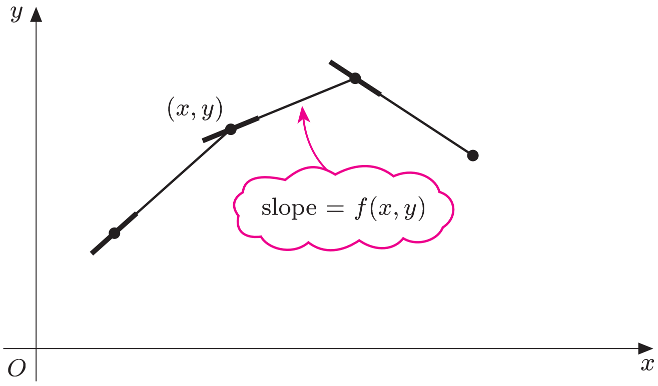


Figure 4.1 Graphical construction for Euler's method, where $dy/dx = f(x, y)$

You will see this idea explained in greater detail in the next activity.

Activity 4.3 Introducing Euler's method

Open Mathcad file **121C3-02 Euler's method**, read the introduction and turn to page 2 of the worksheet. Here a direction field has been set up for the function $f(x, y) = e^{\cos x} - 1$, that is, for the differential equation

$$\frac{dy}{dx} = e^{\cos x} - 1.$$

You studied this differential equation in Activity 4.2(a).

Also, an initial condition is specified, as

$$y = 0 \text{ when } x = 0; \quad \text{that is,} \quad y(0) = 0.$$

In the file, the initial values of x and y are denoted respectively by x_0 and y_0.

Note, to the right of the graph, that the value $f(x_n, y_n) = 1.718$ is displayed, where $n = 0$. This is (to 3 d.p.) the value of

$$f(0, 0) = e^{\cos 0} - 1 = e - 1,$$

which is the slope of the direction field at the starting point $(x_0, y_0) = (0, 0)$. This starting point is denoted on the direction field by a small blue box, and the direction of the direction field at this point is coloured magenta.

You will now see how an approximate solution to this initial-value problem can be built up graphically, step by step. Ensure that all of page 2 from the heading 'Solution curve' to the bottom of the graph is visible on your screen.

(a) Change the value under 'Number of steps' in turn to $N = 1$, $N = 2$, $N = 3$, $N = 4$ and $N = 5$. In each case, observe the effects that the change in value causes.

(b) Observe the effect of changing, in turn, the step size to $h = 0.5$ and the number of steps to $N = 10$.

(c) Observe the effect of changing the initial values. For example, set $x_0 = 1$ and $y_0 = 2$.

Comment

Comment on part (a) for $N = 1$

◇ When 'Number of steps' is changed to $N = 1$, the first segment of the approximate solution curve is drawn, from the starting point $(x_0, y_0) = (0, 0)$ to the point $(x_1, y_1) = (1,\ 1.718)$. The coordinates of these points appear in the tables to the right of the graph. The slopes of the direction field at these points, $f(x_0, y_0) = 1.718$ and $f(x_1, y_1) = 0.717$, are also shown.

All of the values are shown to three decimal places. The slope of the direction field at (x_1, y_1) is now shown by a line segment on the graph, coloured magenta.

The two broken blue line segments (one coinciding with the x-axis) that appear on the graph are temporary construction lines. They illustrate how the first segment of the solution curve is drawn as the hypotenuse of a right-angled triangle. The base (run) of this triangle is equal to the *step size*, which is specified above the graph as $h = 1$. The triangle is constructed, as shown in Figure 4.2, so that *the gradient of its hypotenuse is equal to the gradient given by the direction field at its left-hand vertex*. This gradient is $f(x_0, y_0)$. On the other hand, the gradient (slope) of the hypotenuse is given as usual by the rule 'slope equals rise over run', where the run is the step size, h. The rise, which is the height of the triangle, is therefore given by

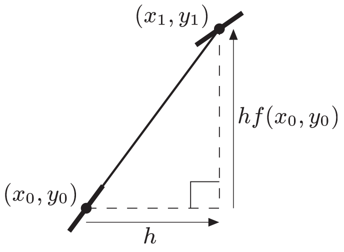


Figure 4.2 First segment

$$\text{rise} = \text{run} \times \text{slope} = hf(x_0, y_0).$$

The coordinates (x_1, y_1) of the top vertex of the triangle can then be calculated from the coordinates (x_0, y_0), the rise and the run. This gives

$$\begin{aligned}(x_1, y_1) &= (x_0 + \text{run},\ y_0 + \text{rise})\\ &= (x_0 + h,\ y_0 + hf(x_0, y_0))\\ &= (0 + 1,\ 0 + 1.718) = (1,\ 1.718).\end{aligned}$$

Comment on part (a) for $N = 2$

◇ When the number of steps is changed to $N = 2$, the second segment of the approximate solution curve is added, from the point $(x_1, y_1) = (1,\ 1.718)$ to the point $(x_2, y_2) = (2, 2.435)$. The broken blue lines now illustrate the 'construction triangle' for this second line segment, as shown also in Figure 4.3. The base (run) of this triangle is again equal to the step size, $h = 1$, but the gradient of its hypotenuse is equal to the gradient given by the direction field at (x_1, y_1). This gradient is $f(x_1, y_1) = 0.717$, as indicated by the table on the right of the screen. The coordinates (x_2, y_2) are therefore given by

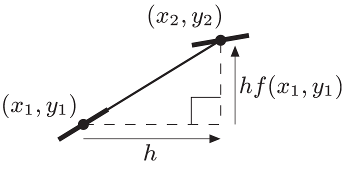


Figure 4.3 Second segment

$$\begin{aligned}(x_2, y_2) &= (x_1 + \text{run},\ y_1 + \text{rise})\\ &= (x_1 + h,\ y_1 + hf(x_1, y_1))\\ &= (1 + 1,\ 1.718 + 0.717) = (2,\ 2.435).\end{aligned}$$

◇ With two steps now completed, it should be fairly clear how the process continues when the number of steps is changed to $N = 3$, $N = 4$ and $N = 5$. The chain of line segments (which form the approximate solution curve) and the rows of corresponding values in the tables to the right of the graph build up one by one as N is incremented. Each time, the latest line segment is joined to the previous one, and its slope matches that of the direction field at its left-hand end.

Comment on part (a) for $N = 3$, $N = 4$ and $N = 5$

Note that the values of $f(x_n, y_n)$ given in the table are negative for $n = 2, 3$ and 4. These correspond to the negative slopes of the direction field at the corresponding points (x_n, y_n), as is clear from the graph.

Mathematically, this means that in constructing the line segment from (x_n, y_n) to (x_{n+1}, y_{n+1}), we have

$$\begin{aligned}(x_{n+1}, y_{n+1}) &= (x_n + \text{run}, y_n + \text{rise})\\ &= (x_n + h, y_n + hf(x_n, y_n)) \quad (n = 0, 1, 2, \ldots).\end{aligned}$$

This is illustrated in Figure 4.4. In other words, the sequences x_n and y_n are determined by the pair of recurrence relations

$$x_{n+1} = x_n + h, \quad y_{n+1} = y_n + hf(x_n, y_n) \quad (n = 0, 1, 2, \ldots),$$

as shown to the right of the graph on screen, together with the specified values for x_0 and y_0. These recurrence relations encapsulate **Euler's method**.

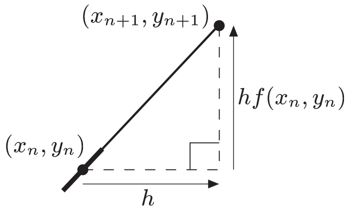


Figure 4.4 $(n + 1)$th segment

Each value y_n is an estimate of the 'true solution' y at $x = x_n$; that is, y_n is an estimate of $y(x_n)$. Clearly, the sequence of estimates obtained depends on the choice of both the step size, h, and the number of steps, N.

◇ With N steps of size h, Euler's method provides $N + 1$ solution estimates, spaced at regular horizontal intervals between the chosen starting value, x_0, and $x_0 + Nh$.

Comment on part (b)

Note that, when starting at $(x_0, y_0) = (0, 0)$, either 5 steps of size 1 or 10 steps of size 0.5 provide a final solution estimate at $x = 5$. By halving the step size h and doubling the number N of steps, their product Nh, and hence also $x_0 + Nh$, remains the same. Notice, however, that the solution estimates obtained at $x = 5$ are different: the final coordinates are $(x_5, y_5) = (5,\ 0.986)$ in the first case and $(x_{10}, y_{10}) = (5,\ 0.558)$ in the second. Halving the step size and doubling the number of steps improves the accuracy of the solution estimate. In the next activity you can investigate further how accuracy improves with reductions in step size.

◇ Setting $x_0 = 1$ and $y_0 = 2$ produces a different chain of line segments, which is again an approximate solution curve for the differential equation, now with the initial condition $y(1) = 2$. (The same is true for other choices of initial values.) If h and N are unaltered from part (b), then the last row of values at the right of the screen corresponds to a point which lies outside the displayed graph region.

Comment on part (c)

Activity 4.3 showed how Euler's method works. You saw a connected chain of line segments built up, one by one, as an approximation to a solution curve. However, this was a fairly inaccurate approximation. The construction utilises slope values provided by the direction field only at points which are a horizontal distance h apart, where the values assigned to the step size h in Activity 4.3 were first 1 and then 0.5. As a result, the approximate solution curve was based on very limited information.

More information can be extracted from the direction field by reducing the step size h, provided that also the number N of steps is increased to maintain coverage of the x-values over which a solution is sought. The next activity shows that such use of extra information leads to improvements in accuracy, and that an estimate for the solution $y(x)$ at a particular chosen value of x can, in principle, be found to whatever accuracy is required.

Activity 4.4 Using Euler's method

You should still be working with Mathcad file 121C3-02.

Turn to page 3 of the worksheet, and read the first paragraph. Then scroll down until all of the page from the heading 'Graphical and numerical solution' to the bottom of the graph is visible on your screen.

The page is set up to solve the same initial-value problem that was considered in Activity 4.3(a) and (b), namely,

$$\frac{dy}{dx} = e^{\cos x} - 1, \quad y(0) = 0.$$

The computation, using Euler's method, takes place from the starting value $x = x_0$ to the finishing value $x = xval$, where $xval$ can be specified by the user. The step size h can be chosen as before. However, in contrast to the situation in Activity 4.3, it is not possible here to vary independently the number of steps, N. Given a step size h, the number of steps to be used is calculated automatically from the condition that the final step is to reach $x = xval$. As a result, the value calculated for y_N is always an estimate for the value $y(xval)$ of the true solution at $x = xval$.

In other words, if h divides exactly into $xval - x_0$, then $N = (xval - x_0)/h$.

(a) Change the value of $xval$ to 9 (from its starting value of 5). What changes on the screen, and what stays the same?

(b) We now seek an estimate for the value of $y(9)$ (the true solution value at $x = 9$). This is obtained by progressively reducing the step size.

First note the value of y_N that appears on screen (this is the estimate for $y(9)$ obtained with step size $h = 1$, and hence with 9 steps).

Now change the step size h in turn to 0.5, 0.2, 0.1, 0.01, 0.001 and 0.0001. In each case, note the corresponding value for y_N. Use these values to estimate the value of $y(9)$ to one decimal place.

Solutions are given on page 86.

Comment

◇ As the step size h is decreased, the number of steps increases. With $h = 0.0001$, there are 90 000 steps, for which the calculation may take an appreciable time on your computer.

◇ With $h = 0.5$, the approximate solution curve still looks like a connected chain of line segments. However, with $h = 0.2$ the graph appears significantly more like a smooth curve, and this remains the case for smaller values of h. (The graph is still in fact made up of short line segments, but so is every graph drawn by Mathcad with the trace type set to 'lines'!)

◇ Just as the numerical estimates for $y(xval)$ appear to converge as the step size is decreased, so too do the approximate solution curves seem to converge towards a 'limiting curve' on the direction field. The graphs obtained provide increasingly close representations of the actual solution curve.

Euler's method, as illustrated in Activities 4.3 and 4.4, was used to plot the solution curves in file 121C3-01, used for Activities 4.1 and 4.2. There, the number of steps was set to $N = 1000$, and calculations were made to draw the line segments to the left, as well as to the right, of the initial point. In that way, the line segments gave the appearance of a smooth solution curve over the whole horizontal graph range.

If you are interested and have the time, then you might like to try the following optional activity, which involves a function $f(x, y)$ that depends on the dependent variable y as well as on the independent variable x. For this particular example, it is also possible to check the outcome against a formula for the solution.

Activity 4.5 Another initial-value problem (Optional)

You should still be working with Mathcad file 121C3-02, on page 3 of the worksheet. Before doing anything else, reset the step size to $h = 1$ (see the first item in the Comment below).

(a) Use Euler's method to estimate to three decimal places the value of $y(6)$, where $y(x)$ satisfies the initial-value problem

$$\frac{dy}{dx} = x - y - 3, \quad y(1) = 1.$$

(For the step size h, use in turn the values 1, 0.5, 0.1, 0.01, 0.001 and 0.0001.)

(b) Check that the function $y = x - 4 + 4e^{1-x}$ is the solution to the initial-value problem in part (a). Hence find $y(6)$ exactly. Does this agree with the value that you obtained using Euler's method in part (a)?

Solutions are given on page 87.

Comment

◇ If you do not start by setting $h = 1$, then any other change will cause recalculation for the most recently-used value of the step size, $h = 0.0001$, which may be time-consuming. Alternatively, recall that any Mathcad calculation can be interrupted by pressing `[Esc]` and then clicking 'OK' in the resulting option box. You may prefer to change here to 'manual calculation mode'.

See the Comment for Activity 4.4 of Chapter B1, in Computer Book B.

◇ There is no need to alter any part of the worksheet before page 3, nor to change the parameters which define the grid for the direction field. On page 3 you need to alter the direction field function definition to $f(x, y) = x - y - 3$. Also, the values of x_0 and y_0 should both be set to 1, and the value of $xval$ to 6.

◇ Using the suggested values of h in turn, the numerical estimates y_N appear to converge and the approximate solution curves do likewise. The 'chain of line segments' is visible for $h = 1$ and for $h = 0.5$, but no departures from smoothness are apparent on the graph for smaller step sizes.

Now close Mathcad file 121C3-02.

Solutions to Activities

Chapter C1

Solution 5.1

Where more than one expression is given below for a solution, the first is similar to the Mathcad output and the second is a form that you are more likely to obtain by hand. (You found each of the derivatives by hand in the main text of Chapter C1, as indicated by the references below.)

(a) $3x^2 - 12x - 15$

See Activity 2.2(a).

(b) $4\pi r^2$

See Activity 3.4(a).

(c) $e^{-t}\left(\cos(t) - 2t\right) - e^{-t}\left(\sin(t) - t^2\right)$

$$= \frac{\cos t - \sin t + t^2 - 2t}{e^t}$$

See Exercise 4.2(b).

(d) $2x\cos(x^2)$

See pages 48–49.

(e) $-4\sin(4x)$

See Activity 4.7(a).

(f) $\dfrac{t^2+3}{t} + 2t\ln(t)$

See Activity 4.2(c).

(g) $\dfrac{1}{u(u^2+3)} - \dfrac{2u\ln(u)}{(u^2+3)^2} = \dfrac{u^2+3-2u^2\ln u}{u(u^2+3)^2}$

See Activity 4.4(b)

(h) $\dfrac{e^t+1}{t+e^t}$

See Exercise 4.3(d).

Solution 5.3

(d) *Step 1:* The one stationary point of the function

$$f(x) = 0.125\sqrt{1+x^2} + 0.0625(2-x),$$

with domain $[0, 2]$, is at $x = 0.577$ (to 3 d.p.).

Step 2: The values of f at the interval endpoints are

$$f(0) = 0.250 \quad\text{and}\quad f(2) = 0.280,$$

while the value of f at the stationary point is

$$f(0.577) = 0.233 \quad \text{(all to 3 d.p.)}.$$

Step 3: Hence the minimum value of $f(x)$ for x in the interval $[0, 2]$ is 0.233 at $x = 0.577$ (both to 3 d.p.).

Hence the solution to the orienteer's problem is to aim to join the path at a distance 0.577 km from the fixed point O on the path.

(This agrees with the answer 0.58 km found in Activity 5.7(a) of Chapter A3 in Computer Book A.)

Solution 5.4

(c) *Step 1:* The stationary points of the function

$$f(v) = \frac{v}{5 + v + 0.02v^2}$$

are at $v = \pm 15.811$ (to 3 d.p.).

Step 2: Only the positive stationary point, 15.811, lies inside the interval $[0, 35]$. The values of f at the interval endpoints are

$$f(0) = 0 \quad\text{and}\quad f(35) = 0.543,$$

while the value of f at the stationary point in the interval is

$$f(15.811) = 0.613 \text{ (all to 3 d.p.)}.$$

Step 3: Hence the maximum value of $f(v)$ for v in the interval $[0, 35]$ is 0.613 at $v = 15.811$ (both to 3 d.p.).

(d) According to the model, a maximum traffic flow rate of about 0.61 vehicles per second can be achieved, by keeping the speed of traffic at about 16 $\mathrm{m\,s^{-1}}$ (that is, about 57 km per hour or 35 mph).

Chapter C2

Solution 5.2

In each case, c is an arbitrary constant which has been added to the expression given by Mathcad. The answers to parts (a) and (b) agree with those obtained earlier by hand.

(a) $\displaystyle\int \left(\frac{1}{x} + e^{3x}\right) dx = \frac{e^{3x}}{3} + \ln(x) + c$

See Activity 1.2(a).

(b) $\displaystyle\int \left(\frac{3}{y^4} + 5\sin(5y)\right) dy = -\frac{y^3\cos(5y) + 1}{y^3} + c$

See Exercise 1.1(b).

(c) $\displaystyle\int (a + \cos(ax))\, dx = \frac{\sin(ax) + a^2x}{a} + c$

Solution 5.3

In each case, c is an arbitrary constant which has been added to the expression given by Mathcad. The Mathcad answers in parts (c) and (d) resemble closely those obtained by hand in the main text.

(a) $\displaystyle\int (x-3)(x-1)\,dx = \frac{x(x-3)^2}{3} + c$

This is equivalent to $\frac{1}{3}x^3 - 2x^2 + 3x + c$. See Example 2.1(a).

(b) $\displaystyle\int \frac{2x-3}{\sqrt{x}}\,dx = -x^{3/2}\left(\frac{6}{x} - \frac{4}{3}\right) + c$

This is equivalent to $\frac{4}{3}x^{3/2} - 6x^{1/2} + c$. See Example 2.1(c).

(c) $\displaystyle\int \sin^2 x\,dx = \frac{x}{2} - \frac{\sin(2x)}{4} + c$

See Activity 2.3(a).

(d) $\displaystyle\int \frac{x}{x^2+1}\,dx = \frac{\ln(x^2+1)}{2} + c$

See Activity 2.5(b)(iii).

(e) $\displaystyle\int u\,e^{3u}\,du = \frac{e^{3u}(3u-1)}{9} + c$

(f) $\displaystyle\int x^2 \ln(5x)\,dx = \frac{x^3\ln(5x)}{3} - \frac{x^3}{9} + c$

(g) $\displaystyle\int \frac{1}{\sqrt{9-t^2}}\,dt = \arcsin\left(\frac{t}{3}\right) + c$

(The function arcsin is represented in Mathcad by 'asin'.)

Solution 5.5

The answers are presented in a form as close as possible to the Mathcad output (which by default is given to 3 decimal places). Each answer agrees with that obtained by hand in the main text.

(a) $\displaystyle\int_0^2 e^t\,dt = e^2 - 1 = 6.389$

See Activity 4.4(b).

(b) $\displaystyle\int_0^{\pi/4} (\cos(5x) + 2\sin(5x))\,dx = \frac{\sqrt{2}}{10} + \frac{2}{5} = 0.541$

See Exercise 4.1(a).

(c) $\displaystyle\int_1^2 \frac{6}{u^2}\,du = 3$

See Exercise 4.1(b). (This answer is the Mathcad outcome from symbolic evaluation.)

(d) $\displaystyle\int_0^{\pi} e^t \sin t\,dt = \frac{e^{\pi}}{2} + \frac{1}{2} = 12.070$

See Exercise 4.1(c). (By default, Mathcad does not show the trailing zero in the third decimal place.)

Solution 5.7

The answers are given by Mathcad to 3 decimal places.

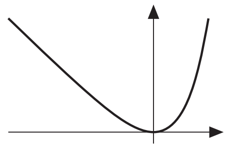
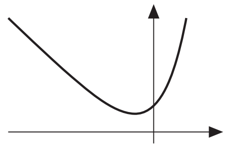
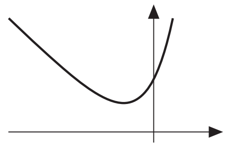
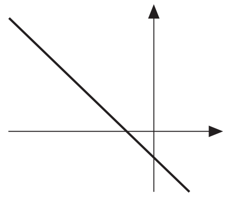
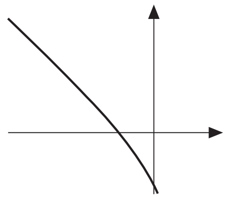
(a) $\displaystyle\int_0^1 t^2\,dt = 0.333$

(b) $\displaystyle\int_{-1}^1 \sqrt{1-x^2}\,dx = 1.571$

Chapter C3

Solution 4.1

(a) The solution curves are described and sketched in the following table.

(x_0, y_0)	Description	Sketch
$(0, 0)$	Rough U-shape; steeper to right of minimum than to left.	
$(0, 1)$	Similar, but with minimum to the left and higher.	
$(0, 2)$	Similar, but with minimum still further to the left and higher.	
$(0, -1)$	Straight line $y = -x - 1$.	
$(0, -2)$	Downward curve with gradient decreasing (i.e. negative but increasing in magnitude).	

(b) The solution curves show three distinct types of behaviour:

(i) Any solution curve through a point above the line $y = -x - 1$ remains above that line. It decreases to a minimum and then increases.

(ii) The line $y = -x - 1$ is itself the solution curve through any point on that line.

(iii) Any solution curve through a point below the line $y = -x - 1$ remains below that line and decreases.

The solution curves through the points $(-3, -1)$, $(-1, 0)$ and $(4, 2)$ are of types (iii), (ii) and (i), respectively.

Solution 4.2

(a) (ii) The slope of the direction field at a point (x_0, y_0) appears to depend on the choice of x_0 alone and not on that of y_0. Correspondingly, there will be solution curves of just one type, with any two such curves differing only by a vertical translation. (This is a consequence of the fact that the function $f(x, y)$ in this case depends only on x.)

(iii) All choices for x_0 and y_0 give a solution curve which increases and decreases alternately, but has a rising trend. Each such curve may be obtained from that shown below by a vertical translation.

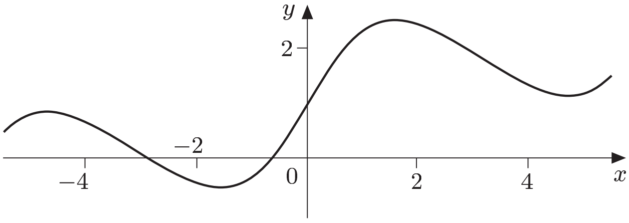


Figure S3.1

(b) (ii) The effect of the direction field on solution curves in this case is not so clear-cut. However, it appears that there may be three types of solution curve, as detailed below.

(iii) There are solution curves that cross the x-axis and are increasing, as in the graph below. If $x_0 = 0$, then such curves are obtained by a choice of y_0 such that $-1.25 \le y_0 \le 1.25$.

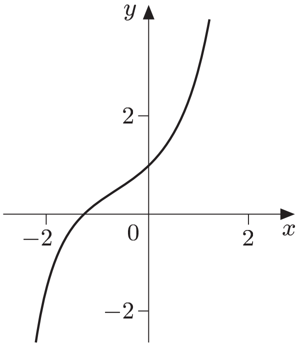


Figure S3.2

There are solution curves that lie above the x-axis, each having a minimum in the second quadrant, as in the graph below. If $x_0 = 0$, then such curves are obtained by a choice of y_0 such that $y_0 \ge 1.26$.

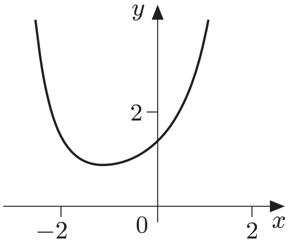


Figure S3.3

There are solution curves that lie below the x-axis, each having a maximum in the fourth quadrant, as in the graph below. If $x_0 = 0$, then such curves are obtained by a choice of y_0 such that $y_0 \le -1.26$.

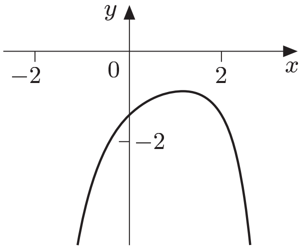


Figure S3.4

Solution 4.4

(a) The graph extends to $x = 9$, but the portion to the left of $x = 5$ remains unchanged. In particular, the step size h does not alter, though the number N of steps does (so that the graph reaches $x = 9$ rather than $x = 5$).

(b) A table of the values obtained is below.

h	1	0.5	0.2	0.1	0.01	0.001	0.0001
y_N	3.916	3.347	3.002	2.887	2.783	2.772	2.771

The values of y_N appear to be converging and, given the level of agreement between the last two estimates, it seems reasonable to deduce from them an estimate for $y(9)$ that is accurate to one decimal place, that is, $y(9) = 2.8$. (In fact, it looks likely that $y(9) = 2.77$ to 2 decimal places.)

Solution 4.5

(a) The table below gives the values obtained using Euler's method, from file 121C3-02.

h	1	0.5	0.1	0.01	0.001	0.0001
y_N	2.000	2.004	2.021	2.026	2.027	2.027

The agreement of the last two estimates suggests that $y(6) = 2.027$ to three decimal places.

(b) For the given function $y = x - 4 + 4e^{1-x}$, we have

$$y(1) = 1 - 4 + 4e^0 = 1,$$

so that the initial condition of the problem in part (a) is satisfied.

The derivative of the given function is

$$\frac{dy}{dx} = 1 - 4e^{1-x},$$

whereas we have

$$\begin{aligned} x - y - 3 &= x - (x - 4 + 4e^{1-x}) - 3 \\ &= 1 - 4e^{1-x}. \end{aligned}$$

Hence the given function also satisfies the differential equation from part (a), and therefore satisfies all the conditions of the initial-value problem.

The value of this solution function at $x = 6$ is

$$y(6) = 6 - 4 + 4e^{1-6} = 2 + 4e^{-5} \simeq 2.027.$$

This value agrees to three decimal places with that obtained using Euler's method.

Computer Book D

Modelling Uncertainty

Guidance notes

This computer book contains those sections of the chapters in Block D which require you to use your computer. Each of those chapters contains instructions as to when you should first refer to particular material in this computer book, so you are advised not to work on the activities here until you have reached the appropriate points in the chapters.

For advice on how each computer session fits into suggested study patterns, refer to the Study guides in the relevant chapters. The statistical software for Block D will have been installed when you installed the MST121 Mathcad files in preparation for Block A. This block does not draw on Mathcad.

About the software

The software for Block D has two components: *Simulations* and *OUStats*. You will be using *Simulations* in Chapters D1 and D3. The main component of the software is *OUStats*. This is a data analysis package, which is introduced in Chapter D2 and used in each of the remaining chapters of the block.

In this block, and in particular when using *OUStats*, it is assumed that you are familiar with the following topics.

Frequency diagrams
Median and quartiles
Boxplots
Mean
Standard deviation
Scatterplots

Frequency diagrams are needed from the start; they are used in Section 1 of Chapter D1. The mean is first mentioned in Section 4 of Chapter D1; the standard deviation and scatterplots are used in Chapter D2 (in Sections 2 and 4, respectively). The median, quartiles and boxplots are used in Chapter D4 after being reviewed briefly in Section 1 of that chapter.

Chapter D1, Section 2
Exploring the problems

The first activity takes you through some of the basic features of the probability simulations which are provided as part of the statistics software.

Activity 2.1 *Probability simulations*

Alternatively, you can access *Simulations* directly by double-clicking on the **MST121 Simulations** icon on your desktop.

(a) To locate the *Simulations* package, click on the **start** menu, move the mouse pointer to **All Programs**, then **MST121**, and finally click on **MST121 Simulations**. After a pause, you should see a 'Home' window featuring the following options.

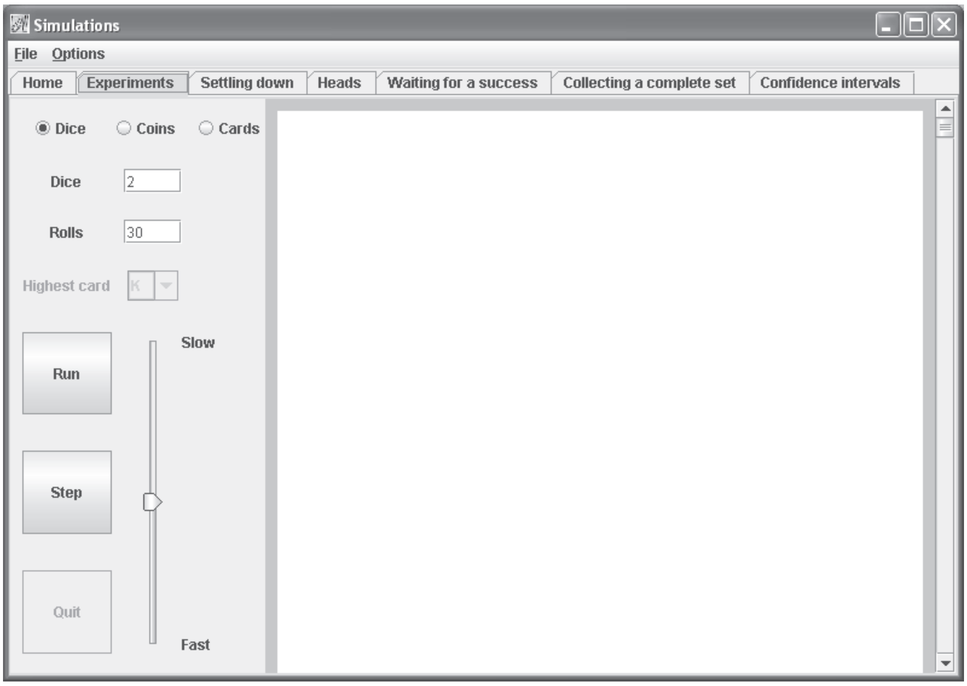
Experiments
Settling down
Heads
Waiting for a success
Collecting a complete set
Confidence intervals

Each of these is accessible via a tab near the top of the window or via a panel towards the right-hand side.

In this section, you will be using all of these simulations except the last; Confidence intervals will be used in Chapter D3.

When you wish to leave the package, click on the **File** menu and then **Exit...** (or simply click on the 'Close' button at the top right-hand corner of the window).

(b) Click on **Experiments** (either the tab or the panel) to open this simulation, and you should see the window shown in Figure 2.1.



● Dice ○ Coins ○ Cards

Dice 2

Rolls 30

Highest card K

Figure 2.1 The opening window for the **Experiments** simulation

Near the top left of the window are three buttons, labelled **Dice**, **Coins** and **Cards**. The default option is **Dice**. A different option can be selected either by clicking on its name (**Coins** or **Cards**) or by clicking on its button.

Select **Coins**.

Now enter the following settings in the boxes near the top left of the window.

Coins 1
Tosses 30

The procedure for doing this is explained below, in case you are not sure how to do it.

You will notice that when the option **Coins** is selected, you are provided with default values for the number of coins and the number of tosses; these values are 2 and 30, respectively. Clicking once in a box positions the cursor in the box. The number in the box can then be edited using the cursor keys, the `[Delete]` key and the `[Backspace]` key. Try typing in a different value for the number of coins now.

You can move the cursor to another box by clicking in the box. Alternatively, if you press the `[Tab]` key, the cursor will move to the next box in the window (in this case, the box for the number of tosses). This number can then be edited as described above. Try doing this now. Then change the settings to 1 coin and 30 tosses.

If you press the `[Tab]` key repeatedly, the cursor selects each of the boxes or buttons within the window in turn. Try this now. You can make use of this feature to run the simulation, or to quit, without using the mouse. For example, if the option marked **Run** is selected in this way, then confirming it by pressing the `[Space]` bar will run the simulation. However, it is usually simpler to run the simulation by clicking on **Run**. Do this now.

Follow the outcomes of 30 tosses of a coin as they appear on the screen. Once the simulation is completed, you can use the scroll-bar (on the right of the window) to scroll back through the outcomes.

(c) Two features of the simulation yet to be mentioned are the speed slider, which is located to the right of the **Run** and **Quit** buttons, and the **Step/Pause** button. By dragging the speed slider with the mouse, the speed of the simulation can be altered. Try running the simulation once more and, while it is running, adjust the speed using the slider. First slow it down, then speed it up.

Dragging a screen object involves placing the mouse pointer on the object and then moving the mouse while holding down the mouse button.

Situated below the **Run** button is a button labelled **Step**. Each time **Step** is clicked on, one toss of a coin is simulated. Click on **Step** several times now to see this effect.

Next, click on **Run** and set the speed to Slow using the speed slider. Notice that the **Step** button is now labelled **Pause**. Click on **Pause**: the simulation is interrupted. You can continue the simulation either by clicking on **Run** or by clicking on **Step**. The effect of clicking on the **Quit** button is to terminate the currently running (or paused) simulation. Spend a few moments exploring these facilities.

If nothing happens when you click on **Dice** or **Cards**, then check that your previous simulation has finished and is not still running slowly or 'paused'.

(d) Now explore the options **Dice** and **Cards**. When you have finished using this simulation (or indeed any of the simulations), you can return to the 'Home' window by clicking on the **Home** tab near the top left of the window. (Alternatively, you can switch directly to one of the other simulations, by clicking on its tab.) When you have finished exploring the options within the **Experiments** simulation, click on **Home**.

You have now explored the first computer simulation. Before going on to use the second one, it is worth reflecting on the art of designing a 'good' computer simulation. On the one hand, a software designer will wish to exploit the power of the computer in order to reduce repetitive routine calculations and tasks, and present only a distilled summary of the reality being simulated. On the other hand, if what you see on the screen differs too much from this reality (the repeated tossing of coins, or whatever), the result can seem abstract and confusing.

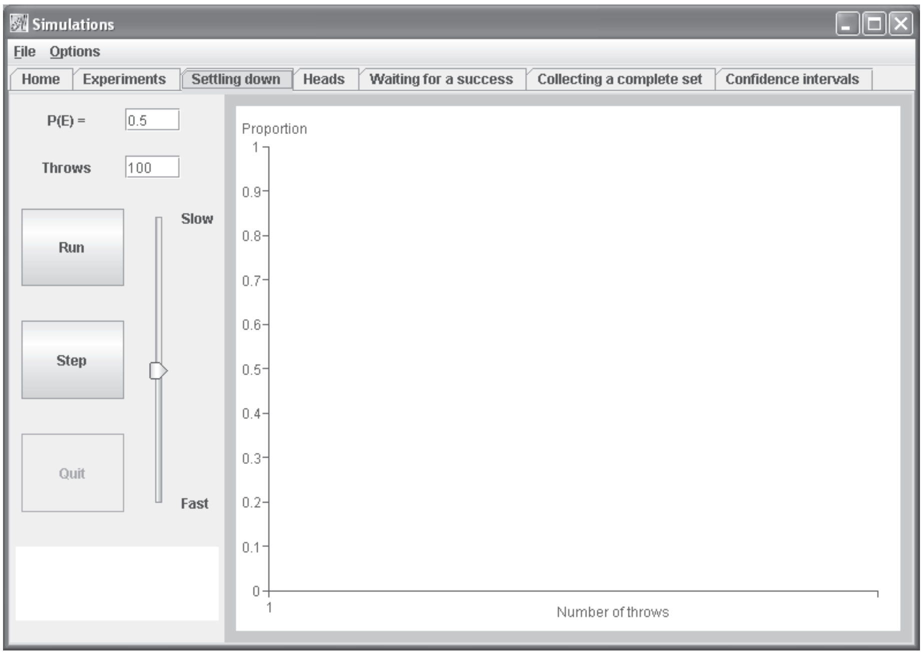
The simulation which you have just run is intended as a sort of halfway house between reality and abstraction. It has the advantage of remaining close to the real-world activity of tossing coins, but it does not exploit the computer fully. The remaining simulations, which are explored in the following activities, exploit the computer more effectively, although the particular situation being simulated may not always be obvious from the screen. As you work through the activities which follow, make sure that you understand what each simulation represents.

Activity 2.2 invites you to use the **Settling down** simulation to explore the 'settling down' phenomenon observed in Activity 1.1 of Chapter D1.

Activity 2.2 Settling down

The command in the **Options** menu allows you to choose **Thick lines...** for this simulation.

(a) Open the **Settling down** simulation and you will see the window shown in Figure 2.2.



P(E) = 0.5

Throws 100

Figure 2.2 The opening window for the **Settling down** simulation

(b) The default number of throws is 100. Reset the number of throws to 30. Now click on **Step** and observe the outcome in the bottom left part of the screen. Notice how the result is displayed on the graph. Click on **Step** several times, checking as you do this how the graph corresponds to the data at each step. The simulation is 'doing' what you did with a real coin in Activity 1.1 of Chapter D1.

The first click on **Step** generates *two* simulated throws.

If you now click on **Run** to complete the simulation, the graph will be completed. Note that at any time you can alter the speed at which the simulation runs (by dragging the speed slider). You can also interrupt the simulation and step through it at any point during its run (by clicking on **Pause** and then on **Step**). Thereafter, you can return to **Run** at any time. Clicking on **Quit** will terminate the simulation.

(c) Now increase the number of throws to 100, and run the simulation once more. Run the simulation for 500 throws and then for 1000 throws. Is the settling down effect apparent in your simulations?

Comment

We cannot predict precisely the results of your simulations. However, it is likely that the settling down effect did become more marked as the number of throws increased.

Activity 2.3 The Brains Trust

Dr Joad defined the law of averages as follows:

> if you spin a coin a hundred times, it will come down heads fifty times, and tails fifty times.

The **Heads** simulation can be used to simulate tossing one or more coins up to 1000 times. The maximum number of coins allowed by the simulation is 20. Use this simulation to investigate the number of heads obtained when one coin is tossed 100 times. Then read the comment below.

Comment

Using the **Heads** simulation, I entered the following settings.

Coins $\boxed{1}$

Tosses $\boxed{100}$

Throughout this computer book, 'I' refers to a particular member of the MST121 module team who carried out these activities. The results reported have no special status, but can be used to provide a counterpoint to what you found, as well as enabling discussion of specific points that arise from the particular data obtained.

Then I ran the simulation.

There are two buttons, **Frequencies** and **Proportions**, near the top left of the window. When **Frequencies** is selected, the frequency is displayed at the top of each bar on the graph. When **Proportions** is selected, proportions are displayed. (To select an option, click on its button.)

For my simulation, the frequency displayed on the '1 head' bar was 44. When I selected **Proportions**, the proportion of tosses which resulted in a head was displayed: this was 0.44.

I ran the simulation a further nine times, and each time noted the frequency on the '1 head' bar. The ten frequencies were as follows.

44 49 51 48 55 46 53 46 45 52

None of my ten simulations produced exactly 50 heads, which appears to knock Dr Joad's definition firmly on the head! However, the average number of heads obtained in these ten runs is 48.9, which is quite close to 50. So the average proportion of tosses which resulted in a head was close to $\frac{1}{2}$.

Perhaps Dr Joad's definition of the law of averages could be reworded as follows:

> if you spin a coin a large number of times, the proportion of spins that result in a head will be approximately $\frac{1}{2}$.

Activity 2.4 D'Alembert's heads

D'Alembert argued that, in two tosses of a coin, there are three possible outcomes – heads on the first toss, heads on the second toss, and heads on neither toss. By his reasoning, since two of these three give at least one head, the probability that the coin lands heads at least once is $\frac{2}{3}$.

Remember that, for our purposes, tossing two coins is equivalent to tossing one coin twice.

(a) Use the **Heads** simulation to investigate d'Alembert's conclusion. This time two coins are being tossed, so set the number of coins to 2. You may need to run the simulation several times, with various different numbers of tosses, in order to reach a conclusion. Record your results in a table like the one below, and write down your conclusions.

Run	Number of tosses	Tosses which gave at least 1 head	
		Number	Proportion
1	100		
2			
3			
⋮			

(b) Do you think that d'Alembert's conjectured probability of $\frac{2}{3}$ is correct? If not, having carried out some simulations, what do you think the correct value of the probability is? Do the results of your simulations agree with the ideas you jotted down in Subsection 1.3 of Chapter D1?

Comment

We shall return to this problem in Section 3 of the main text.

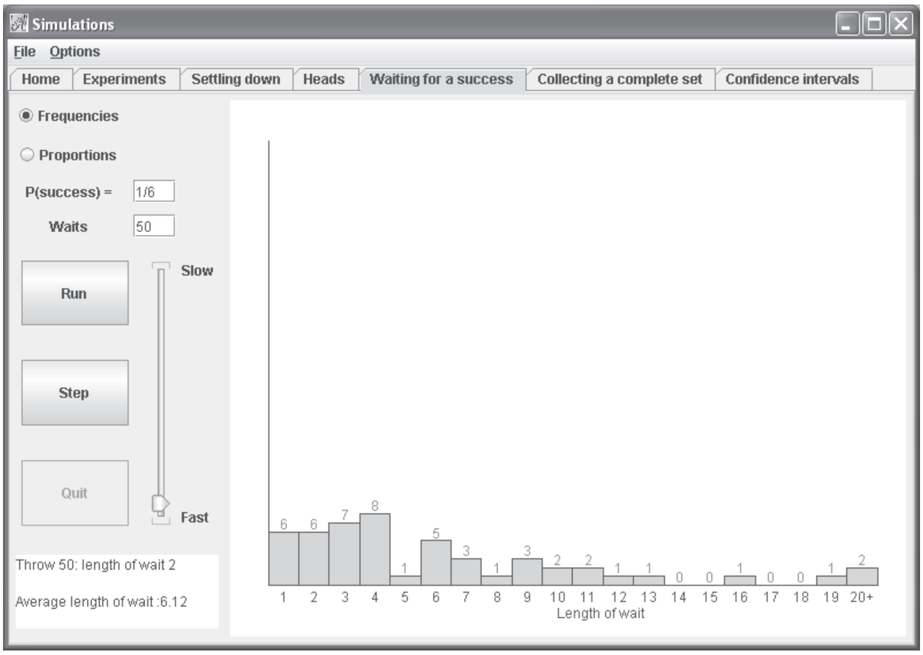
Activity 2.5 Waiting for a six

In some board games, players can join in only when they roll a six with a die. In Subsection 1.3, you were invited to write down your ideas concerning several questions about the length of time (measured as the number of rolls of a die) that a player has to wait to join in a game. The **Waiting for a success** simulation can be used to investigate these questions.

(a) Open the **Waiting for a success** simulation. Each time a die is rolled, the probability of obtaining a six is $\frac{1}{6}$. If we regard obtaining a six as a success, then $P(\text{success}) = \frac{1}{6}$. The number of times the die has to be rolled to obtain a six (a success) is the wait. On the screen, enter the following values for the settings: $P(\text{success}) = 1/6$ and 1 wait.

Run the simulation several times, and try to get a sense of what lengths of wait, typically, tend to occur.

(b) Now set the number of waits to 50, and step through the first few waits to ensure that you understand what is going on. Then click on **Run** to complete the simulation. You should obtain output similar to that shown in Figure 2.3.



P(success) = 1/6

Waits 50

Throw 50: length of wait 2

Average length of wait :6.12

Figure 2.3 The results of a run of the **Waiting for a success** simulation

Notice that if, as in the simulation depicted in Figure 2.3, you obtained some waits that were longer than 19, then these are registered on the '20+' bar at the right-hand end of the horizontal axis.

Notice also the extra information that appears at the end of a simulation in the box in the bottom left part of the window. This is the average length of the waits in the simulation just run.

(c) Run the simulation several times. Can you tell from your simulation when you are most likely to achieve a six? That is, what number of rolls is most likely to be needed to obtain a six?

Comment

Overall, you may have found that your results were very variable and it was difficult to draw any firm conclusions based on just 50 waits. A greater number of waits is clearly necessary. You are asked to try this in the next activity.

Activity 2.6 Still waiting for a six

In this activity, you are invited to continue your investigation from Activity 2.5 by running the simulation a number of times for a larger number of waits. As you do so, focus on the following questions.

◇ On average, how many times will a player have to roll a die in order to obtain a six?

◇ What is the most likely number of rolls needed to obtain a six?

You may find it helpful to record your results in a table like the one below.

Number of waits	Average wait	Most likely number of rolls
300		
300		
300		
⋮		

(a) Run the simulation several times using 300 waits. On each occasion, note the average length of the waits and the number of rolls (that is, the length of wait) which occurred most frequently. What do you notice about the frequencies of the different wait lengths? How would you describe the general shape of the frequency diagram?

(b) Use your results to form hypotheses about the answers to the two questions above. Experiment with different numbers of waits to help you do this.

(c) How do your hypotheses compare with your intuitions? Were you surprised by any of the results that you obtained?

Comment

The problem *Waiting for a six* is investigated further in Section 4 of the main text.

Activity 2.7 How long is an average wait?

The **Waiting for a success** simulation can be used to investigate the waiting time for other events; for example, the number of tosses of a coin needed to obtain a head, or the number of children a couple might need to have to produce a girl. If we assume that $P(\text{head}) = \frac{1}{2}$ and $P(\text{girl}) = \frac{1}{2}$, then in both examples we can use the simulation with $P(\text{success}) = \frac{1}{2}$. Other events would require different values of $P(\text{success})$.

(a) Use the simulation to explore the average wait for various values of $P(\text{success})$. Note down your results, and use them to predict a value for the average wait for each value of $P(\text{success})$ that you choose. You may find it helpful to record your results in a table like the following one.

$P(\text{success})$	Number of waits	Average wait: observed values	Average wait: prediction
$\frac{1}{6}$			
$\frac{1}{2}$			
$\frac{1}{5}$			
0.4			
⋮			

(b) Can you spot any pattern in your results? If $P(\text{success}) = p$, what would your conjecture be for the average wait?

Comment

We shall return to this problem in Section 4 of the main text.

The final two computer activities in this section use the **Collecting a complete set** simulation. It has been designed to allow you to investigate the problem *Collecting a complete set of musicians.*

Activity 2.8 Collecting a complete set of musicians

One out of eight different toy musicians is given away in each packet of a popular breakfast cereal. In Subsection 1.3 of Chapter D1, you were asked to guess the number of packets of cereal that you might expect to have to buy in order to collect a complete set of eight musicians. In this activity, you are invited to investigate this problem using the **Collecting a complete set** simulation.

(a) After opening the window for this simulation, change the number of objects in a set to 8 (and leave the number of collections at 1). To make sure that you understand what this simulation does, step through the simulation until you obtain a complete set. At each step, each object has an equal chance of being selected. Objects are selected until at least one of each different type has been chosen. The number of the last object selected is highlighted on the horizontal axis and recorded in the box at the bottom left of the window. The number of packets needed to complete the collection is eventually displayed in this box also.

If you run the simulation for a number of collections greater than 1 then, when the simulation finishes, all of the results are recorded in the box at the bottom left of the window. The number of packets needed to complete each collection is displayed after 'Packets:'. If the results are not all visible, then they can be viewed by scrolling through them. If you click on a line in this box, representing a particular collection, then the corresponding diagram is displayed.

(b) Now run the simulation several times, each time noting the number of packets required to obtain a complete set. Run the simulation so as to obtain at least 10 collections, and write down the number of packets that were required for each collection. (You will need to refer to your results again in Section 5 of the main text.)

Number of packets required to collect a complete set											

(c) Did you find your results surprising? Are they consistent with the ideas you noted in Subsection 1.3? Do you wish to revise the conjecture you made in Subsection 1.3 for the average number of packets required to collect a complete set? If so, then make a note of your revised prediction.

Comment

What you might have found striking was the great variability in the number of packets needed to collect a complete set of eight musicians. In Section 5 of the main text, we return to the problem of finding the average number of packets required.

Activity 2.9 Collecting a complete set

The **Collecting a complete set** simulation can be used to explore the numbers of packets needed to collect complete sets of sizes other than 8. Consider the following situations.

(a) Take a well-shuffled pack of cards, and cut the pack at random, noting the card revealed. Repeat this procedure until you have selected at least one card from each suit. How many times would you expect to have to cut the pack?

(b) You toss a coin repeatedly, noting the result (head or tail) each time. You stop at the point when you first have at least one occurrence of each outcome. How many tosses would you expect to have to make?

Explore each of these questions using the **Collecting a complete set** simulation.

Chapter D2, Section 3
Fitting a normal model

3.1 Introducing OUStats

In this subsection, you will be guided through some of the facilities of *OUStats*, the data analysis part of the statistics software.

Nearly all the data sets that you will be exploring and analysing in this chapter (and in the rest of Block D) are large, and simply typing the data into *OUStats* would occupy a lot of time. In consequence, all the data sets are provided as data files. You will not be asked to enter data sets yourself, either in this block or in assignments, so you will not find any instructions here for entering or editing data. However, instructions have been included in an Appendix in case you wish to be able to use *OUStats* to analyse your own data (or in case you are interested to know how this is done). Whether or not you study the Appendix is entirely up to you.

Start up *OUStats* now, as follows.

◇ Click on the **start** menu, move the mouse pointer to **All Programs**, then **MST121**, and finally click on **MST121 OUStats**. The following window should appear.

Alternatively, you can access *OUStats* directly by double-clicking on the **MST121 OUStats** icon on your desktop.

Instructions for ‘clicking’ here and below refer to use of the left-hand mouse button, unless stated otherwise.

The underlined letters in menu titles indicate the keystroke with `[Alt]` that can be used (instead of a mouse click) to open the menu. Thus `[Alt]F` will open the **File** menu.

Note that you can exit from *OUStats* at any time, simply by clicking on **File** and choosing **Exit** (by clicking on it). Alternatively, click on the ‘Close’ button at the top right-hand corner of the window.

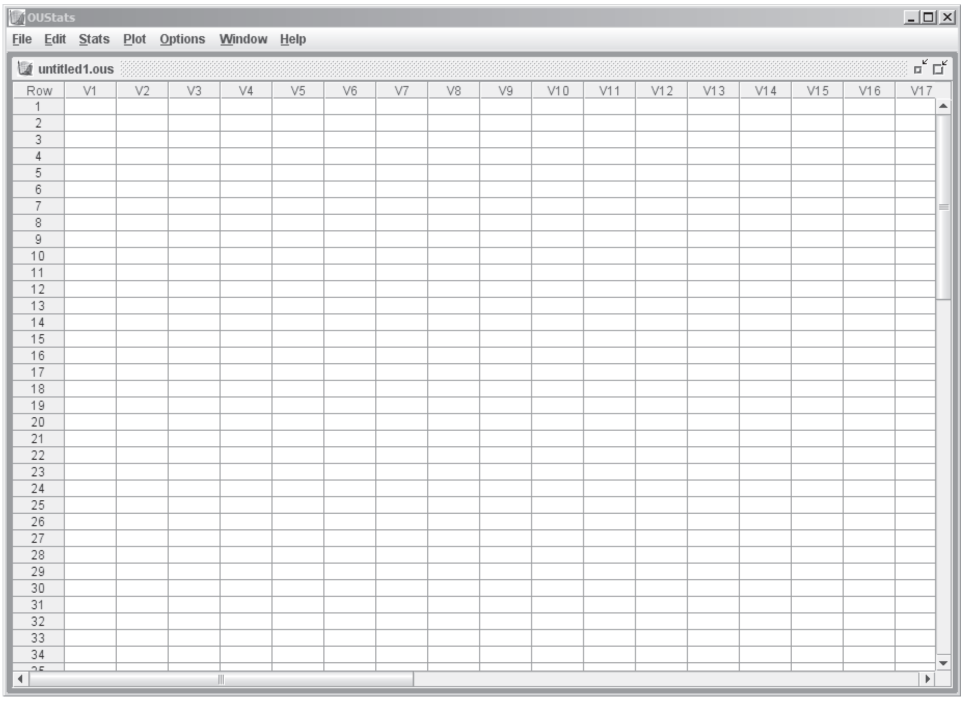


Figure 3.1 The opening window in *OUStats*

This is a blank data window, with title ‘untitled1.ous’. It is made up of cells, with rows numbered 1, 2, 3, … and columns labelled V1, V2, V3, … .

Across the top of the window are seven menu titles: **File**, **Edit**, **Stats**, **Plot**, **Options**, **Window** and **Help**. Four of these (**File**, **Edit**, **Window**, **Help**) should look familiar from other *Windows*-based packages.

◇ Click on **File**.

Alternatively, use the keystroke combination `[Alt]F`.

As you can see, this menu contains commands for handling files.

◇ Click on **Edit**.

The commands in this menu allow you to edit the data window, rename columns, and so on.

◇ Click on **Window**.

The first two commands here, **Cascade** and **Tile**, provide different rearrangements of the windows that you create.

◇ Click on **Help**.

The first command gives access to the on-screen Help facility.

The other menus, **Stats**, **Plot**, **Options**, are specific to this software package.

◇ Click on **Stats**.

The commands here provide for a range of calculations to be carried out on data. To use most of them, you need first to open or create a data file. You will be asked to open such a file shortly, and the use of some of these commands will then be demonstrated. Other commands will be used in later chapters: **Confidence interval...** in Chapter D3, **Two sample z-test...** and **Regression...** in Chapter D4.

◇ Click on **Plot**.

This menu contains the commands for obtaining diagrams to represent data. You will see use of **Frequency diagram...** shortly; **Scatterplot...** will be used later in this chapter, and **Boxplot...** in Chapter D4.

Bar chart... and **Pie chart...** are not used in MST121, but you may wish to explore their use at another time.

◇ Click on **Options**.

The first two commands here allow you to change some of the settings of the package, and in particular the number of significant figures to which output values are displayed.

This completes a first look at the options available from the seven menus. You are next asked to open a data file.

◇ Click on **File**. Then choose **Open...** (by clicking on it).

All of the file names shown have the extension '.OUS', indicating that they are *OUStats* data files.

◇ Scroll through the list until you find the name GEYSER.OUS. Click on this name (adding it to the **File Name** box), then click on **Open**.

You should now have a data window that contains a single column of data, headed 'Intervals'. When a data file is open, you can obtain information about the data using **Notes...** in the **File** menu.

◇ Click on **File**, and select **Notes...**. Read what this text says about the data in GEYSER.OUS. Then remove the 'Notes' window by clicking on the 'Close' button at its top right-hand corner.

You will next see how to obtain summary statistics and a frequency diagram for this data set. This will show how intervals between eruptions of the geyser vary.

◇ Click on **Stats**, then select **Summary stats...**. In the window that appears, select 'Intervals' (by clicking on it) and click on **Select**.

After a possible pause, summary statistics for the data set appear lower in the window. These include the mean, standard deviation, minimum value, and so on. The last item is the sample size, which is 299. Also shown are the median and quartiles; these will be reviewed in Chapter D4. Note (for reference shortly) that the minimum and maximum values are, respectively, 43 and 108 (in minutes). Next we obtain a frequency diagram.

◇ Click on **Plot**, then select **Frequency diagram...** . From the left-hand drop-down menu, select 'Intervals' (which is the only item in this case). Then click on **Go**.

The frequency diagram that appears in the window uses a first interval starting value and interval width that have been generated automatically by the software. These values can also be specified by the user. We noted earlier that the data had minimum value 43 and maximum value 108. These values suggest that a first interval starting value of 40 and interval width of 10 might be suitable.

◇ Click on the **First interval** box. Edit the contents so that 'auto' is replaced by 40. Then edit the **Width** box so that 'auto' is replaced by 10. Finally, click on **Go**.

The second box can be reached either by clicking on it or by using the `[Tab]` key.

The frequency diagram is redrawn using these values. You now have three open windows: the original data window, a 'Summary statistics' window and a 'Frequency diagram' window. It is possible to see all of these together, in an orderly format, using the **Tile** facility.

◇ Click on **Window**, then select **Tile**.

The three windows now each occupy a quarter of the overall *OUStats* window, with a blank space in the remaining quarter.

This concludes our look at output arising from the data in GEYSER.OUS. In order to close the file, it suffices to open another one, which will be used to demonstrate further features of *OUStats*.

◇ Click on **File**. Then choose **Open...** . Locate the file name HEIGHTS.OUS and click on it, then click on **Open**.

The file HEIGHTS.OUS contains frequency data.

◇ Click on **File**, and select **Notes...** . Read what this text says about the data in HEIGHTS.OUS.

First we obtain summary statistics for these data.

◇ Click on **Stats**, then select **Summary stats...** .

The list of variables here contains three items, 'Height', 'Frequency' and 'Height | Frequency'. The first two just correspond to the values held in the columns of the same names, but 'Height | Frequency' takes account of the linkage between the two columns (that is, that a height of 62 inches occurs 3 times, 63 inches occurs 20 times, and so on). Hence it is the last choice which is the correct one here.

◇ Choose 'Height | Frequency', then click on **Select**.

The sample size is 1000 (Cambridge men). The values range from a minimum of 62 inches to a maximum of 77 inches. (Note these two values for reference shortly.) Now we turn to a frequency diagram.

◇ Click on **Plot**, then select **Frequency diagram...** . From the left-hand drop-down menu, select 'Height | Frequency'. Set **First interval** to 60 (slightly less than the minimum of 62 noted above) and set **Width** to 2. Finally, click on **Go**.

The shape of the frequency diagram (which is the same shape as that of the corresponding histogram) indicates that a normal model might be suitable for the heights of Cambridge men. So next we fit a normal curve to the data.

You saw this also in Section 1.

◇ In the second drop-down menu from the left within the 'Frequency diagram' window, select 'Fit normal curve'.

The display changes from a frequency diagram to a histogram. Also new boxes appear, indicating the mean (68.872) and standard deviation (2.567 92) of the data set. If the **Fit normal curve** button is clicked on, then a normal curve with the indicated mean and standard deviation is added to the histogram. However, it is also possible to edit the values in the **Mean:** and **Std dev:** boxes, so that they have more appropriate accuracy (three significant figures, say).

Note these numbers for use shortly.

◇ Change the value in the **Mean:** box to 68.9. Then change the value in the **Std dev:** box to 2.57. Finally, click on the **Fit normal curve** button.

The chosen normal curve is now superimposed on the histogram. It does indeed seem to fit the data well. The curve is a model for the heights of all Cambridge men in 1902. It can be used to estimate, for example, the proportion of Cambridge men in that year who were between 69.5 and 70.5 inches in height. To find this proportion, we seek the area under the curve between 69.5 and 70.5. This can be found using **Normal distribution...** from the **Stats** menu.

◇ Click on **Stats**, then select **Normal distribution...** . Within the window that appears, change the value in the **Mean** box to 68.9, and (after pressing `[Tab]` to change boxes) change the value in the **Standard deviation** box to 2.57 (both as noted above). Then click on **Update** to update the graph below.

The normal curve shown corresponds to the mean and standard deviation just entered. Above the curve (and below the **Update** button) are four boxes and two buttons. The first two boxes, **A** and **B**, show their preset values (which are both 0). The two lower boxes show the corresponding values of **Area to left of A** and **Area between A and B** (which are currently both displayed as 0). The window can now be used in either of two ways:

The area to the left of A is actually non-zero but extremely small, since A = 0 is many standard deviations away from the mean.

(a) to input values for A, B and find the corresponding areas;

(b) to input values for the areas and find the corresponding values for A and B.

To demonstrate the first of these, we find the area under the curve between 69.5 and 70.5.

◇ Click on the **A** box (or use `[Tab]` repeatedly to reach this by navigation around the boxes and buttons). Change the value in this box to 69.5. Press `[Tab]` to move to the **B** box, and change the value here to 70.5. Now click on **Use A & B to calc areas**.

The area between A = 69.5 and B = 70.5 is now shown shaded on the graph. The value of this area is displayed in the **Area between A and B** box (and also on the graph). The value of the area under the curve to the left of A is also shown both in a box and towards the left on the graph.

The value of the shaded area is 0.141 (to three significant figures). This means that, according to the model, 14.1% of Cambridge men in 1902 were between 69.5 and 70.5 inches tall. Another conclusion here, with reference to the **Area to left of A** box, is that, according to the model, 59.2% of Cambridge men in 1902 were less than 69.5 inches tall.

Put another way, if a man had been chosen at random from this population, then the probability that his height would have been between 69.5 and 70.5 inches is 0.141.

We now use the window in the second way described above, by finding a value of A that depends on a given value of the area to the left of A. Suppose that we seek the height such that 95% of men in this population were shorter than this height.

◇ Click on the **Area to left of A** box and enter 0.95 into it. Then click on **Use areas to calc A & B**.

An error message appears, indicating that the sum of the two areas (that to the left of A, and that between A and B) must be less than 1.

This is as it should be, since the total area beneath the normal curve is 1.

◇ Click on **OK**, and enter in the **Area between A and B** box any non-negative number less than 0.05 (0 suffices). Then click again on **Use areas to calc A & B**.

Alternatively, areas can be selected by clicking and dragging with the mouse on the graph. See later for further details.

The value obtained in the **A** box is 73.1273. Hence, according to the model, 95% of Cambridge men in 1902 were less than about 73.1 inches tall.

That concludes this introductory tour of some *OUStats* facilities. You may wish to spend some time on your own at this point to explore other features of the package. For reference purposes, the main facilities of *OUStats* introduced so far are summarised below.

(1) The menus

The **File** and **Edit** menus contain commands for handling and editing files.

The **Stats** menu contains commands for calculations.

The **Plot** menu contains commands for obtaining diagrams.

The **Options** menu allows you to change some of the settings of *OUStats*. These include the number of significant figures shown in displayed results and the colours to be used in displayed graphs. You can also alter a 'graphics smoothing' setting in order to improve the appearance of printouts obtained from graph windows. Additionally, the Calculator resident in *Windows* is accessible via this menu.

The **Window** menu contains commands for arranging windows on the screen. (These are standard *Windows* operations.)

The **Help** menu provides access to on-screen help.

(2) Data files

To open a data file:

◇ click on **File** and choose **Open...** (by clicking on it) – a dialogue box appears;

◇ click on the file name, then click on **Open**.

Information about the data in the file currently open can be obtained by choosing **Notes...** in the **File** menu.

(3) Summary statistics

To obtain summary statistics for a variable in an open data file:

◇ click on **Stats**, then choose **Summary stats...** (by clicking on it);

◇ select the variable name(s) from the scrolling list in the new window that appears (by clicking on it/them), then click on **Select**.

To select more than one variable at a time, hold down the `[Ctrl]` key while making your choice.

The summary statistics are then displayed in the window.

(4) Frequency diagrams

To obtain a frequency diagram:

◇ click on **Plot**, then choose **Frequency diagram...** (by clicking on it);

◇ select the variable name from the left-hand drop-down menu in the window that appears;

◇ enter the first interval starting value in the **First interval** box and the interval width in the **Width** box, should you not want the software to choose these values automatically, then click on **Go**.

Each such diagram produced is displayed in a separate window, titled 'Frequency diagram:' followed by the variable name.

Histograms are produced in a similar way. Choose the second drop-down menu from the left within the 'Frequency diagram' window, and select 'Histogram' in place of 'Frequency'.

(5) Fitting a normal curve to data

To fit a normal curve:

◇ obtain a frequency diagram for the data, as above;

◇ in the second drop-down menu from the left, select 'Fit normal curve' (the display then changes from a frequency diagram to a histogram);

◇ enter the mean and standard deviation of the required normal curve in the appropriate boxes (or accept the values provided), then click on the **Fit normal curve** button.

The default values are the sample mean and sample standard deviation.

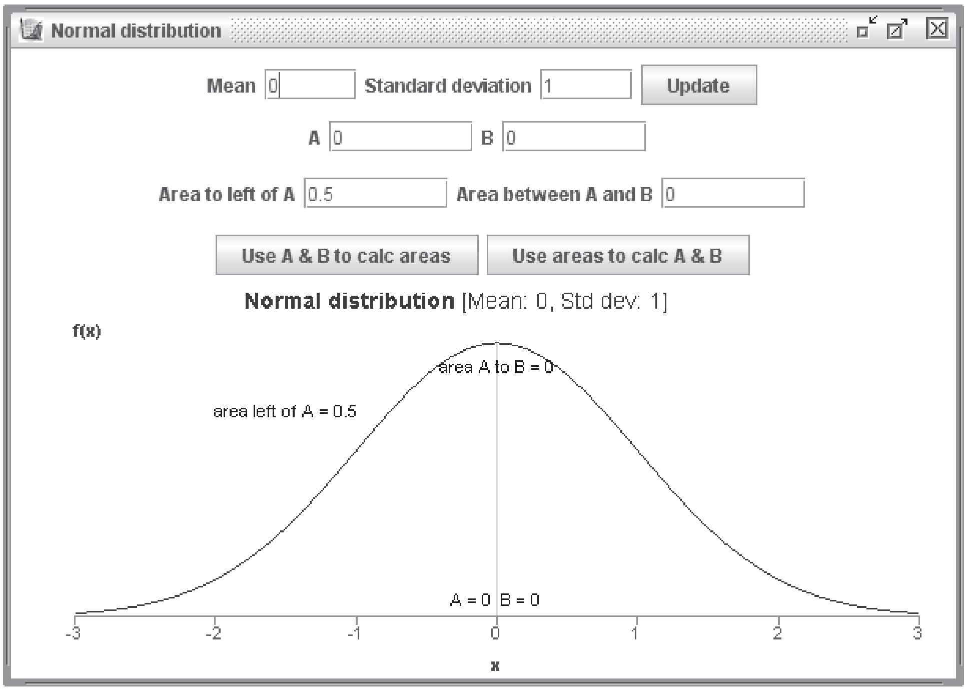
The normal curve with the specified mean and standard deviation is then superimposed on the histogram.

(6) Finding areas under a normal curve

To find an area under a normal curve, you must first:

◇ click on **Stats**, then choose **Normal distribution...** (by clicking on it), whereupon a window similar to that in Figure 3.2 on the next page will appear;

◇ enter the mean and standard deviation of the required normal distribution in the appropriate boxes, and click on **Update**.



Normal distribution

Mean 0 Standard deviation 1

A 0 B 0

Area to left of A 0.5 Area between A and B 0

Figure 3.2 The opening window for **Normal distribution...**

To find the area to the left of a numerical value, a (say):

- *either* edit the value in the **A** box to read a and the value in the **B** box to be greater than a, then click on the **Use A & B to calc areas** button;
- *or* position the mouse pointer close to the horizontal axis, hold the mouse button down and drag the mouse until the value in the **A** box is as close as possible to a, then let go of the mouse button.

You may position the mouse pointer anywhere in the window, outside of boxes and buttons, but you will probably find it more useful to position it close to the horizontal axis.

The area to the left of a will appear in the **Area to left of A** box (and also above the graph).

To find the area between two values, a and b (say):

- enter the value a in the **A** box and the value b in the **B** box, by either of the methods just described.

Dragging the mouse from left to right increases the value of B (from a starting value for A), while dragging from right to left decreases the value of A.

If you choose to press and drag the mouse, then the area under the curve between a and b is shaded as soon as you let go of the mouse button, and the area is given in the **Area between A and B** box (and also above the graph). If you edit the values, then the area is shaded and its value is given as soon as you click on the **Use A & B to calc areas** button.

To find the value a such that the area to the left of a is equal to a numerical value p (say):

- edit the **Area to left of A** box so that it contains the number p;
- click on the **Use areas to calc A & B** button.

The value of p must be less than 1. Also, the value in the **Area between A and B** box should be less than $1 - p$.

The value a will be displayed in the **A** box.

(7) Saving files and copying from windows

If you make any changes to one of the data files supplied with *OUStats*, and wish to save the amended data window, then you must do so using **Save As...** from the **File** menu, and you must save it using a different name: all the data files and notes files supplied are protected so that you cannot accidentally change the contents of the original files.

The computer operating system does not distinguish between upper- and lower-case letters in file names: you can use either. For clarity of presentation, we have used upper-case letters in file names in this computer book.

Active output from an *OUStats* window can be copied to other applications in the usual way, using Copy and Paste.

Any window other than the original data window can be closed using the 'Close' button at its top right-hand corner. The data window can be closed only by: (a) opening another data file; (b) creating a new data file; (c) exiting from the application.

When a data file is opened, any previously-saved files are closed down automatically. You will receive a prompt inviting you to save the data window if it has been changed. No windows other than the data window and accompanying Notes can be saved.

(8) Using frequency data

Some data are stored in two columns in a data file, with values in one column and corresponding frequencies in the next. When using such data – to obtain summary statistics or a frequency diagram, for instance – note that the two column names appear on a single line in variable lists, with a vertical bar between them: for example, 'Value | Frequency'. However, the individual column labels (in this case 'Value' and 'Frequency') also appear in variable lists as potential choices, so it is important to remember to choose the option with the vertical bar where appropriate.

A comment on frequency diagrams

Before you move on to the next group of computer activities in Subsection 3.2, there is one point concerning frequency diagrams that ought to be mentioned. When using a computer package, it is all too easy to obtain a frequency diagram without giving much thought to whether the diagram you obtain is the 'best' possible. Different choices of starting values and interval widths will, in general, produce different diagrams representing the same data; and not all choices of the starting value and interval width will necessarily be appropriate for the data.

Consider, for instance, the two frequency diagrams for the heights of 1000 Cambridge men that you produced earlier. The first was obtained by allowing *OUStats* to select automatically the starting value of the first interval and the width of the intervals. For the second, we chose these values ourselves: since the heights ranged from 62 to 77, it seemed reasonable to choose 60 as the starting value and 2 as the interval width. But were these good choices?

You may recall that the heights were recorded to the nearest inch. (This information is given in the Notes that accompany HEIGHTS.OUS and was mentioned when the data were introduced in Section 1.) This means that, for instance, the heights of men who were anywhere between 61.5 and 62.5 inches tall were recorded as 62 inches, the heights of men between 62.5 and 63.5 inches tall were recorded as 63 inches, and so on.

For the second frequency diagram, we specified that the first interval should start at 60 and have width 2. So *OUStats* included in this interval all heights recorded as at least 60 inches but less than 62 inches; there were none, since the lowest recorded height was 62 inches. The second interval included all heights recorded as at least 62 inches but less than 64 inches, that is, all those recorded as either 62 inches or 63 inches; there were 23 of these. So there were 23 men between 61.5 and 63.5 inches tall. The heights of these men were represented on the frequency diagram by a bar drawn from 62 to 64, when clearly a bar drawn from 61.5 to 63.5 would have been better.

This sort of discrepancy can be avoided by noting how the data were recorded and choosing intervals appropriately. In this case, since the shortest recorded height was 62 inches and heights between 61.5 and 62.5 inches were recorded as 62, it would be sensible to choose 61.5 as the starting value of the first interval, rather than 60 or 62. Then the first bar would be drawn from 61.5 to 63.5. By the way, the frequency diagram in Figure 1.2 of Chapter D2 can be obtained by using a starting value of 61.5 for the first interval and an interval width of 1.

The important message to be obtained from this example is that when you use a statistics software package, you need to think about what you are doing. Although a package will save you all the work involved in doing calculations and drawing diagrams, it will not think for you. It will usually do whatever you ask it to, whether or not your instructions are sensible or appropriate.

3.2 Is a normal model a good fit?

In Subsection 3.1, a normal curve was chosen to model the variation in the heights of Cambridge men in 1902, but no check was made on whether the model was a 'good' fit. In this subsection, we use *OUStats* to investigate informally, for several samples of data, whether a fitted normal curve is a good model for the variation observed in the data.

The first stage in investigating whether a normal distribution is a suitable model should always be to obtain a frequency diagram for the data, and to inspect its shape. If it is clearly not bell-shaped – for example, if it is skewed or has more than one clear peak, such as for the four frequency diagrams in Figure 1.4 of Chapter D2 – then a normal model can be rejected immediately. But if it looks as though a bell-shaped curve might be a suitable model for the variation in the data (as in Figure 1.5), then the next step is to fit a normal curve with parameters μ, estimated by the sample mean $\overline{x}$, and σ, estimated by the sample standard deviation s. The fit of the curve can then be inspected by eye.

Sometimes (as was the case for the heights of Cambridge men) a histogram for the data and the fitted normal curve are so similar in shape that it is clear that the model is a good one. But more commonly, perhaps because of the jaggedness of the histogram, there is some doubt about the fit. The problem is to decide whether the differences between the histogram and the curve could be the result of chance, and just a feature of the particular sample, or whether they are an indication that the model is not a good fit.

There are formal statistical tests that can be carried out, called goodness-of-fit tests, for deciding whether a chosen model is a good one for the variation in a sample of data; and if you study statistics in the future, then you will almost certainly meet such tests. However, in MST121, we adopt a more informal approach, using simulations to generate samples from the chosen normal distribution. This gives an indication of the nature of the variation that occurs by chance, and thus offers a benchmark against which to judge whether the data could reasonably be thought to be a sample from the chosen normal distribution.

In the following activities, you are invited to explore whether a normal model is a good fit for each of a number of data sets. In the first activity, you are asked to decide for each data set, by looking at a frequency diagram, whether a normal curve is even worth considering. In each of the subsequent activities, you are asked to fit a normal model, and then use simulations to investigate the suitability of the model.

Activity 3.1 Is a normal model worth considering?

In this activity, for each data set, you should obtain a frequency diagram for the data and hence decide whether or not a normal distribution might be suitable for modelling the observed variation. For each data set, you should go through the following steps:

◇ open the data file;

◇ read the information about the data contained in **Notes...** ;

◇ obtain a frequency diagram for the data, taking particular care over your choice of the first interval starting value and interval width;

◇ decide whether or not a normal distribution might be suitable for modelling the observed variation, explaining your decision briefly.

Instructions are given for the first data set only.

(a) The file DIPPER.OUS contains the weights in grams of 198 Irish dipper nestlings at age 6–8 days. Open the data file now. (Choose **Open...** from the **File** menu, click on DIPPER.OUS, and then click on **Open**.) Read the information on these data contained in **Notes...** . (Choose **Notes...** from the **File** menu.)

You may need to scroll through the list of file names to find the file.

The weights are grouped – the individual weights are not given. The groups are listed in the first column of the data window, the midpoints of the groups are in the second column (labelled 'Weight'), and the frequencies are in the third column. To represent these data sensibly on a frequency diagram, you need to start the first interval at 9 and use 2 as the interval width.

Obtain a frequency diagram for the data using these values. (Choose **Frequency diagram...** from the **Plot** menu, click on 'Weight | Frequency' in the left-hand drop-down menu, enter 9 and 2 as the starting value for the first interval and the interval width, respectively, and finally click on **Go**.)

Now consider whether a normal distribution might be suitable for modelling the variation in the data. Questions you might consider include the following. Is the frequency diagram roughly bell-shaped, or is it skewed? Does it have more than one peak?

(b) Repeat the process in part (a) for the data on radial velocities of stars which are contained in the data file RADIAL.OUS.

(c) Repeat the process in part (a) for the data on the lengths of sentences written by H. G. Wells which are contained in the data file AUTHORS.OUS.

(d) Repeat the process in part (a) for the data on the lengths of cuckoo eggs which are contained in the data file CUCKOOS.OUS.

Solutions are given on page 144.

In each of the next three activities, you should:

◇ fit a normal curve to the data;

◇ generate random samples from the fitted normal distribution;

◇ compare frequency diagrams for the random samples with a frequency diagram for the data.

Fairly detailed instructions are given in the first activity below. You should follow a similar procedure for the other two.

Activity 3.2 Weights of Irish dipper nestlings

(a) *Fitting a normal curve*

Open the data file DIPPER.OUS, and obtain a frequency diagram for the data using a first interval starting value of 9 and an interval width of 2 (as you did in Activity 3.1).

From the second drop-down menu from the left, select 'Fit normal curve'. The display changes from a frequency diagram to a histogram. New boxes appear, indicating the mean (27.9394) and standard deviation (7.747 04) of the data set, displayed to six significant figures by default.

These statistics are chosen because they are estimates for the population mean and population standard deviation; but it would not be reasonable to suppose or claim six-figure accuracy for these estimates. *As a rough guide, it is reasonable to quote sample statistics to one significant figure more than is given in the data used to calculate them.* The weights of the Irish dipper nestlings are given to two significant figures (roughly), so three-significant-figure accuracy is appropriate for the sample mean and sample standard deviation.

It is a good idea to jot down the parameters of the normal curve that you fit, as you will need them again later.

Enter the values 27.9 and 7.75 for the mean and standard deviation of the normal curve. Click on the **Fit normal curve** button, and the normal curve is fitted over the histogram.

It looks as though a normal curve fits the data quite well, although the frequency diagram is quite jagged. To see whether this is the sort of jaggedness that might be expected to occur by chance, we shall compare this data set with some random samples *of the same size* drawn from the normal distribution that we have fitted to the data. The sample size is important here: because of the 'settling down' effect, we should expect small samples to produce more jagged frequency diagrams than large ones, so we must compare the frequency diagrams of random samples of the same size as the sample of data.

(b) *Obtaining random samples from the fitted normal distribution*

Random samples from a normal distribution are obtained using **Normal samples...** from the **Stats** menu. Click on **Stats** and choose **Normal samples...** ; a dialogue box appears. Enter 27.9 for the mean and 7.75 for the standard deviation; these are the parameters of the normal curve that you just fitted. The sample size should be the same as for the data, 198 in this case, so enter 198 in the **Samples** box. To generate three random samples, enter 3 in the **Sets** box. Finally, click on **OK**.

The samples are generated and stored in the first three empty columns in the data window: these columns are labelled 'Random1', 'Random2' and 'Random3'.

(c) *Comparing the random samples and the data*

We want to compare the frequency diagram of the data with frequency diagrams for the random samples. We shall use the same intervals for all the samples. Before obtaining the frequency diagrams, we need to decide on the starting value for the first interval; and to do this we need to know the lowest value that appears in any of the samples. This could be found by looking at the data and picking out the lowest value, but this is a tedious exercise for large sample sizes. An alternative is to use **Summary stats...** , since the minimum is one of the statistics displayed.

Choose **Summary stats...** from the **Stats** menu. You can select several variables at the same time by holding down the `[Ctrl]` key. Do this and click on 'Random1', 'Random2' and 'Random3' in turn. Then release the `[Ctrl]` key. Click on **Select**, and summary statistics for all three samples will be displayed lower in the window. Pick out the minimum value for each random sample. (You will need to scroll through the output to see them all.)

The minimum value in one of my samples was 5.725 66. This was the lowest of the three minimum values, so I decided to use 5 as the starting value for the first interval for the frequency diagrams (instead of 9, which was used for the data earlier). Your random samples will be different from mine, so you may need to use a different starting value. Remember that we want intervals 9–11, 11–13, and so on, so the starting value of the first interval must be an odd number.

Now we have all the information that we need. For each of the variables 'Weight | Frequency', 'Random1', 'Random2' and 'Random3', proceed as follows. Choose **Frequency diagram...** from the **Plot** menu, and select the variable name from the left-hand drop-down menu in the window. Input your starting value for the first interval and 2 for the interval width, then click on **Go**. Having done this for each of the variables, you should have created four frequency diagrams.

(d) *Viewing the frequency diagrams*

Use **Tile** from the **Window** menu so that you can see all of the open windows at once. Close down any open window other than the data window and the four frequency diagrams just created. Maximise the size of the overall *OUStats* window. Then use **Tile** again, to see the four frequency diagrams together and at similar size. My four diagrams are shown in Figure 3.3.

The data window cannot be closed.

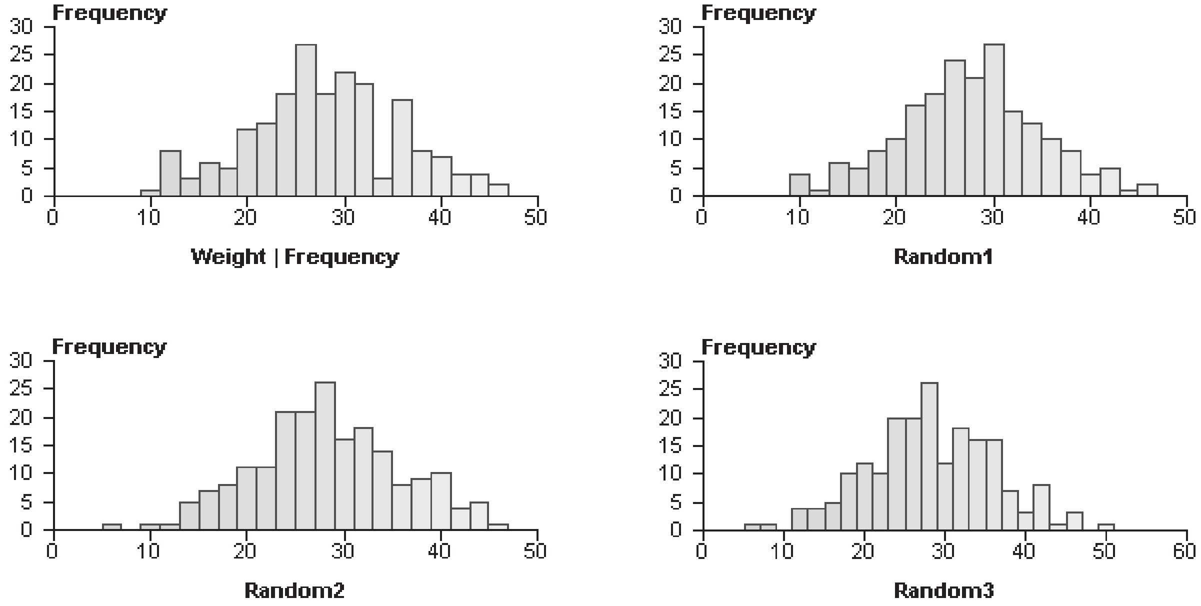


Figure 3.3 Four frequency diagrams from *OUStats*

We are now in a position to compare the variation in the random samples from the normal distribution that we fitted with that in the data, and hence to decide informally whether or not the model is a good fit. As you can see, the variability evident in the frequency diagrams for the random samples is broadly similar to that in the frequency diagram for the data. So it looks as though the normal distribution is indeed a good model for the variation in the weights of Irish dipper nestlings aged 6–8 days.

Your random samples will be different from mine. Are your frequency diagrams similar in shape to those obtained here? Notice that the vertical scales on all these frequency diagrams are the same. However, this may not have been so for your diagrams: the scales depend on the values in the samples being represented, and three of the samples were random samples. When comparing frequency diagrams, you need to note where corresponding scales are different from each other, even within windows of the same overall size. It is possible to change the size of a window, in order to make equal distances on two scales appear to have the same length on screen. Usually this is not necessary for the comparison of frequency diagrams, but since it is a facility which can also be applied usefully in other circumstances, the box below explains how it can be done.

Note that the horizontal scale of the diagram for 'Random3' in Figure 3.3 differs from that of the other three diagrams.

Numerical scales within a resized *OUStats* window continue to feature the same numbers, but the diagram is squashed or stretched along with the window.

Changing the size of a window using the mouse

When the mouse pointer is positioned on the border of a window, a double-headed arrow across the border replaces the mouse pointer on the screen. To make a window wider or narrower, place the mouse pointer on either the left or right edge of the window so that this double arrow is showing. Then press down the mouse button and drag the mouse sideways. When you release the mouse button, the window will be redrawn with the edge in the new position. Similarly, you can make a window taller or shorter by pressing and dragging while the mouse pointer is positioned on either the top or the bottom edge of the window. Any diagram inside the window is resized to fit the new window.

Activity 3.3 Radial velocities

Repeat the steps described in Activity 3.2 to investigate whether a normal model is a good fit for the variation observed in the radial velocities which are contained in the data file RADIAL.OUS.

Some comments are given on page 144.

Activity 3.4 Lengths of cuckoo eggs

Repeat the steps described in Activity 3.2 to investigate whether a normal model is a good fit for the variation observed in the lengths of cuckoo eggs which are contained in the data file CUCKOOS.OUS.

Some comments are given on page 145.

Generating random samples: a summary

Random samples from a normal distribution are obtained as follows.

◇ Choose **Normal samples...** from the **Stats** menu.

◇ Enter the mean and standard deviation of the normal distribution, the sample size (in the **Samples** box) and the number of random samples required (in the **Sets** box), then click on **OK**.

If k samples are generated, then they are stored in the first k available columns in the data window and are named 'Random1', 'Random2', . . . , 'Random$k$'. (If further random samples are generated at a later stage, possibly with a different data window but during the same *OUStats* session, then the numbering in 'Random' labels continues from the last number used previously.)

Normal samples and random numbers

The command **Normal samples...** , which is contained in the **Stats** menu of *OUStats*, allows you to generate samples of values chosen randomly from a normal distribution. You may have wondered how this is done.

It may seem paradoxical to use a computer to produce 'random' numbers: we expect any computer program to produce output that is entirely predictable. Nevertheless, computer 'random number generators' are in common use; these generate sequences of 'random' integers. Given an initial value – the *seed* value – the sequence generated is predictable and therefore it is not truly random. Numbers generated in this way are called *pseudo-random* numbers. However, in practice, sequences of pseudo-random numbers are indistinguishable from sequences of random numbers, so they may be regarded as sequences of random numbers and used to simulate random samples in statistical simulations.

Most computer programming languages have a routine that generates pseudo-random integers between zero and the maximum integer N that can be stored by the computer. These integers can be used to generate 'random' values from any distribution – normal, geometric or whatever. The details of how this is done are beyond the scope of MST121.

The MST121 statistics software uses your computer's clock to determine the seed value of the underlying random number generator, thus ensuring that each statistical experiment (using *Simulations*) or random sample is different. For example, each time you use the **Normal samples...** command in *OUStats*, the seed value is set using the current date and time, making it extremely unlikely that the samples you obtain are the same as any you have obtained previously.

Note that for lotteries, premium bond draws, etc., random simulations based on a pre-programmed algorithm are not used, as they could be open to discovery. Instead, some physical randomising device is used. One such device is based on the number of electrons moving inside a valve.

3.3 Printing with OUStats

The main steps involved in printing with *OUStats* are set out below.

First, make sure that your printer is connected, is installed under *Windows*, and is switched on. The instructions depend in part on whether you wish to print text windows (containing only text) or graph windows.

To print text windows from *OUStats*:

◇ activate each window that you want to print by clicking on it (or by choosing its title from the **Window** menu);

◇ choose **Print...** from the **File** menu, and click on **OK**.

The output from the window will then be printed.

To print graph windows from *OUStats*:

◇ prior to creating the graph windows, choose **Graph options...** from the **Options** menu and make sure that the **Use graphics smoothing** box is *unchecked*;

◇ create the windows to be printed, then proceed as above for printing text windows.

The checking of this box gives a clearer image when viewed on screen but a less distinct image when printed on paper.

Alternatively, if a graph window is created before the **Use graphics smoothing** box is unchecked, then after unchecking this box you will need to right-click on the window to obtain the **Refresh** menu, and then to click on this to alter the graphics quality. Following this, proceed as above for printing text windows.

You may like to use the following activity to check that you can print output from *OUStats*.

Activity 3.5 Printing output from OUStats

Open the file DIPPER.OUS.

Calculate summary statistics for the weights of the Irish dipper nestlings; these are displayed in a new window.

Now obtain a frequency diagram for the data; this is displayed in another new window.

Now print the output from each of the two windows that have been produced, following the instructions for printing given above.

Chapter D2, Section 4
Are people getting taller?

In the second half of the 19th century, considerable interest developed in the inheritance of characteristics, both in plants and in animals and humans. In the 1890s, Karl Pearson (1857–1936) determined to obtain data on three physical measurements – height, span of arms and length of left forearm – for a large number of families. Many of the data were collected by college students, some of whom made measurements on as many as twenty families. The data were collated by Dr Alice Lee, a colleague of Pearson's at University College, London; she calculated various statistics and prepared some 78 tables of data. According to Pearson, 'this occupied her spare time for nearly two years'. In 1903, several of these tables were published in an article in the journal *Biometrika*.

K. Pearson and A. Lee, 'On the laws of inheritance in man', *Biometrika* 2 (1903) pages 357–462.

The box opposite is a verbatim extract from the instructions given to those who collected the data.

The instruction sheet also contained diagrams illustrating the second and third measurements described. The data cards on which the measurements were recorded emphasised that 'both father and mother are absolutely necessary and should not be over 65 years of age' and that neither parent should be a step-parent. All measurements were recorded to the nearest quarter of an inch, although the heights were rounded to the nearest inch before tabulation. A great deal of thought went into the instructions and the design of the data cards. For example, experiments were carried out into the effect of wearing boots on measured heights, and as a result it was decided to subtract an inch from the recorded height of each boot-wearer. As well as noting the wearing of boots, collectors were also asked to put L, A or C against all the measurements if a person being measured had ever broken a leg, arm or collar-bone.

All those measured were between 18 and 65 years old. Pearson explained his choice of this restriction on age in the article. He observed that full growth may not be reached until age 25 or thereabouts. However, he realised that insisting that all sons and daughters should be over 25 years old might make collecting the data much more problematic, not least because it might be difficult to interest college students in the project as most of them were aged between 19 and 22.

There was also the fact that fewer families with all the sons and daughters over 25 years old had both parents surviving. So, since growth between 18 and 25 is very small, he fixed on 18 years as the lower age limit. He also observed that, because of the phenomenon of shrinkage with age, it would have been better to take a lower maximum age than 65 years for parents, but this too would have limited the number of available families.

Altogether, over a thousand families were measured. The heights of 1078 father–son pairs and 1375 mother–daughter pairs were included in the results.

MAYBE, A HUNDRED YEARS FROM NOW, SOMEONE WILL FIND MY LITTLE PROJECT OF INTEREST ?

FAMILY MEASUREMENTS

Professor KARL PEARSON, of University College, London, would esteem it a great favour if any persons in a position to do so, would assist him by making one set (or if possible several sets) of anthropometric measurements on their own family, or on families with whom they are acquainted. The measurements are to be made use of for testing theories of heredity, no names, except that of the recorder, are required, but the Professor trusts to the *bona fides* of each recorder to send only correct results.

Each family should consist of a father, mother, and at least one son or daughter, not necessarily the eldest. The sons or daughters are to be at least 18 years of age, and measurements are to be made on not more than two sons and two daughters of the same family. If more than two sons or daughters are easily accessible, then not the tallest but the eldest of those accessible should be selected.

To be of real service the whole series ought to contain 1000–2000 families, and therefore the Professor will be only too grateful if anyone will undertake several families for him.

The measurements required in the case of each individual are to be to the nearest quarter of an inch, and to consist of the following.

(I.) *Height* – This measurement should be taken, if possible, with the person in stockings, if she or he is in boots it should be noted. The height is most easily measured by pressing a book with its pages in a *vertical plane* on the top of the head while the individual stands against a wall.

(II.) *Span of Arms* – Greatest possible distance between the tip of one middle finger and the tip of the other middle finger, the individual standing upright against a wall with the feet well apart and the arms outstretched – if possible with one finger against a doorpost or corner of the room.

(III.) *The Length of LEFT Forearm* – The arm being bent *as much as possible* is laid upon a table, with the hand flattened and pressed firmly against the table, a box, book, or other hard object is placed on its edge so as to touch the bony projection of the elbow, another so as to touch the tip of the middle finger. Care must be taken that the books are both perpendicular to the edge of the table. The distance between the books is measured with a tape.

Or,
The arm being bent *as much as possible* the elbow is pressed against the corner of a room or the doorpost, the hand being flattened and pressed against the wall. The greatest distance from the tip of the middle finger to the corner or doorpost is to be measured.

In this section, you will have the opportunity to explore the data that Pearson obtained on the heights of father–son pairs. One question we shall investigate is: ‘Were the sons taller, on average, than the fathers?’ That is, was the phenomenon of increasing stature, which has been observed more recently, evident in these families at the beginning of the 20th century? Another question is: ‘Did tall fathers tend to have tall sons, and short fathers have short sons?’ In this section, you will be able to investigate both these questions using the statistics software. We shall return to the second question in Chapter D4.

Activity 4.1 The data

Click on **Open...** in the **File** menu. The data on father–son heights are contained in the file PEARSON.OUS. Open this data file now.

As you can see, the data are arranged in a frequency table, with the columns containing father's height in inches, son's height in inches, and frequency, respectively. You will see that, for example, there was one father–son pair with the father's height recorded as 59 inches and the son's height as 64 inches; and, if you scroll down to row 23, you will see that there were four father–son pairs with the father's height 63 inches and the son's height 67 inches.

For paired data such as these, it is a good idea to begin by obtaining a scatterplot of the data. Click on **Plot** and choose **Scatterplot...** (by clicking on it). A 'Scatterplot' window appears. To obtain a scatterplot with father's height on the x-axis and son's height on the y-axis, select FatherHt | Frequency for the x variable and SonHt | Frequency for the y variable. The scatterplot is then displayed lower in the window.

Is there any pattern discernible in the scatterplot? What does this tell you about the heights of the fathers and the sons?

Comment

The scatterplot is shown in Figure 4.1. (After choosing **Scatterplot...**, the size of the window was adjusted in order to obtain Figure 4.1.)

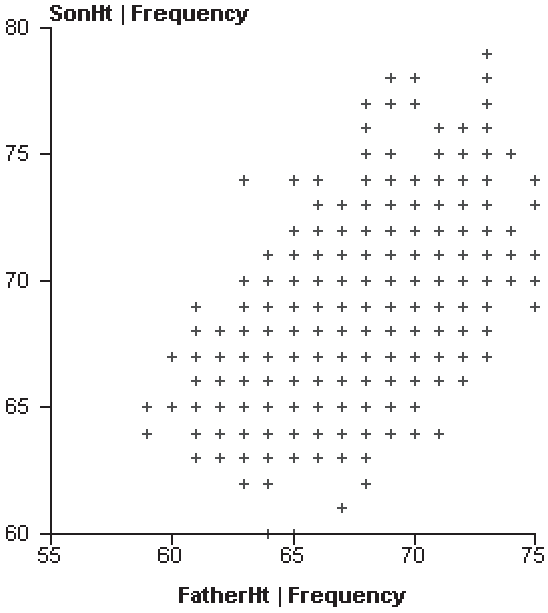


Figure 4.1 A scatterplot of son's height against father's height

Notice that some information is not shown in this scatterplot: it does not show how many father–son pairs there were for each pair of heights, only whether or not there were any pairs. From the scatterplot, it appears that there is a tendency for the taller fathers to have sons taller than those of the shorter fathers. However, there is a lot of scatter, so the relationship between son's height and father's height is a weak one. It is not possible to tell from the scatterplot whether or not the average height of the sons is greater than the average height of the fathers.

Activity 4.2 Average heights

Recall that you can select more than one variable at a time by holding down the `[Ctrl]` key while you click on variable names with the mouse.

Use **Summary stats...** in the **Stats** menu to find the mean and standard deviation of the fathers' heights and of the sons' heights. (Select the variables FatherHt | Frequency and SonHt | Frequency to summarise the fathers' heights and sons' heights, respectively.)

Is the average height of the sons greater than the average height of the fathers? Are the sons' heights and the fathers' heights equally variable?

Comment

The sons were taller on average than the fathers – the mean height of the sons is 68.7 inches compared with 67.7 inches for the fathers. The heights of the fathers and the sons were equally variable – the standard deviations (2.75 inches for the sons and 2.72 inches for the fathers) are approximately equal.

Activity 4.3 Modelling the heights

(a) A frequency diagram of the sons' heights is shown in Figure 4.2(a). A normal curve with mean 68.7 and standard deviation 2.75 has been superimposed on the corresponding histogram in Figure 4.2(b).

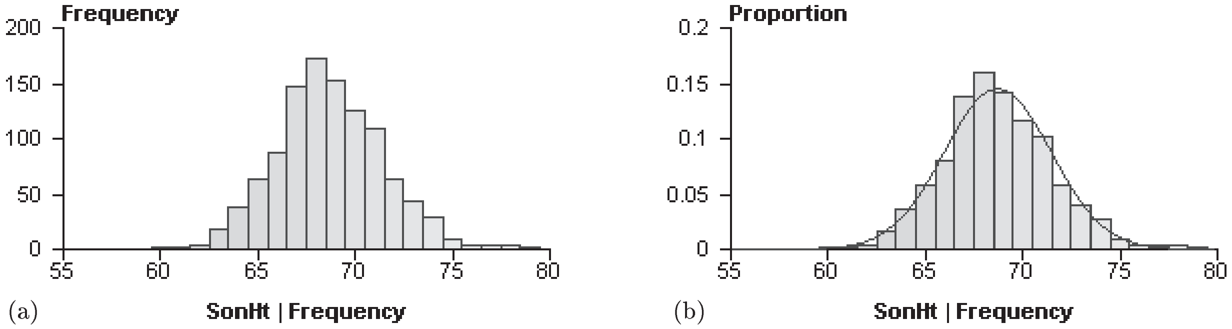


Figure 4.2 (a) A frequency diagram (b) The fitted normal curve

It looks as though a normal distribution models the variation in heights quite well.

Now follow the instructions below for using **Normal distribution...** to find the proportion of sons in this generation who were, according to this model, over six feet tall. This proportion is given by the area under the normal curve to the right of 72; this is shown in Figure 4.3.

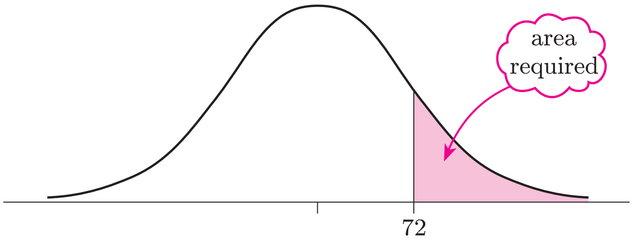


Figure 4.3 The area required

Select **Normal distribution...** from the **Stats** menu, and enter 68.7 and 2.75 for the mean and standard deviation, respectively. Click on **Update**, and the normal curve will be displayed.

First, find the area to the left of 72 as follows. Enter 72 in the **A** box and 72 (or any greater value) in the **B** box. Then click on the **Use A & B to calc areas** button (or `[Tab]` to this button and press the `[Space]` bar). The area is displayed in the **Area to left of A** box: it is 0.885 (to 3 significant figures). Since the total area under any normal curve is 1, the area required is equal to

$$1 - 0.885 = 0.115.$$

So, according to the model, approximately 11.5% of sons in this generation were over six feet tall.

(b) Now fit a normal distribution with mean 67.7 and standard deviation 2.72 to the data on the heights of the fathers. Use this model to estimate the proportion of fathers in this generation who were over six feet tall. Was there a greater proportion of sons over six feet tall than of fathers?

Comment

According to the model for fathers' heights, approximately 5.7% of fathers were over six feet tall. So a greater proportion of sons than fathers were over six feet tall.

In Activity 4.2, you found that the sons were taller, on average, than the fathers. And if anyone whose height is over six feet is regarded as tall, then it appears that there were more 'tall' sons than 'tall' fathers. However, to tackle the question of whether sons are taller than their fathers, we really need to look at the heights of sons whose fathers are of particular heights. For example, is the average height of the sons of fathers who were 64 inches tall greater than 64 inches? And is the average height of the sons of six-foot-tall fathers over six feet? You are asked to investigate questions such as these in the next two activities. The data are stored in a convenient form in the file SONS.OUS, so use this file for these activities. (The data in SONS.OUS and PEARSON.OUS are exactly the same; they are just arranged differently.)

Activity 4.4 Modelling sons' heights

In Activity 4.1, it was observed that sons of tall fathers tended to be taller than sons of short fathers. But if we know the height of a father, what precisely can we say about the height of his son? In this activity, you are invited to explore the heights of sons of fathers of various different heights.

(a) The data on the heights of sons of fathers who were 69 inches tall are contained in the columns named SonHt69 and Freq69 in the file SONS.OUS. Obtain a frequency diagram for these data.

You should find that it looks as though a normal distribution might provide a reasonable model for the variation in these heights. So fit a normal model to these data. (Remember to use one significant figure more for the parameters of the normal distribution than are given in the data: in this case, this means estimating the mean height to one decimal place and the standard deviation to two decimal places.)

According to the model, what proportion of the sons of 69-inch-tall fathers were more than 69 inches tall (and so taller than their fathers)?

(b) Investigate the heights of sons of fathers who were 71 inches tall. (The data are in the columns SonHt71 and Freq71.) Fit a normal distribution to the heights, and use it to estimate the proportion of sons of 71-inch-tall fathers who were taller than their fathers.

(c) Investigate the heights of sons of fathers who were 67 inches tall and of sons of fathers who were 64 inches tall. (These data are in the pairs of columns SonHt67, Freq67 and SonHt64, Freq64, respectively.) In each case, use a normal distribution to estimate the proportion of sons who were taller than their fathers.

Comment

(a) Using a normal distribution with mean 69.5 and standard deviation 2.30 in **Normal distribution...**, I found that the area to the left of 69 was equal to 0.414. So the proportion of sons over 69 inches tall was, according to the model, $1 - 0.414 = 0.586$, or approximately 59%.

(b) Using a normal distribution with mean 70.4 and standard deviation 2.48, I found that the area to the left of 71 was equal to 0.596. So the proportion of sons over 71 inches tall was, according to the model, $1 - 0.596 = 0.404$, or approximately 40%.

(c) Using a normal distribution with mean 68.0 and standard deviation 2.21, I found that the area to the left of 67 was equal to 0.325. So the proportion of sons over 67 inches tall was, according to the model, $1 - 0.325 = 0.675$, or approximately 68%.

Using a normal distribution with mean 66.6 and standard deviation 2.17, I found that the area to the left of 64 was equal to 0.115. So the proportion of sons over 64 inches tall was, according to the model, $1 - 0.115 = 0.885$, or approximately 89%.

It looks as though the sons of short men were more likely to be taller than their fathers than were the sons of tall men.

Activity 4.5 Looking for a relationship

In Activity 4.4, you found that, according to the models, although more than half the sons of fathers 64 inches, 67 inches or 69 inches tall were taller than their fathers, fewer than half the sons of fathers 71 inches tall were taller than their fathers. In this activity you are asked to investigate how the mean height of sons varies with father's height.

During Activity 4.4, you found some mean heights of sons for fathers of particular heights, as shown below.

Table 4.1

Father's height in inches	Mean height of sons in inches
64	66.6
67	68.0
69	69.5
71	70.4

The values in Table 4.1 were obtained from **Fit normal model**, but could also have been found from **Summary stats...** . The table can be extended by finding the mean height of sons in a similar way for every value of father's height from 59 to 75 inches. These data are contained in the file MEANS.OUS. Open this file now and check that the four data pairs in Table 4.1 agree with the corresponding rows of the data matrix.

You can see from the data in the columns FatherHt and SonMeanHt that the mean height of sons tends to increase with father's height. However, the sons of short fathers were not, on average, as short as their fathers; and the sons of tall fathers were not, on average, as tall as their fathers.

Obtain from the file MEANS.OUS a scatterplot, with father's height (FatherHt) on the x-axis and mean son's height (SonMeanHt) on the y-axis. What does the scatterplot tell you about the relationship between the heights of fathers and the mean height of their sons?

Comment

The scatterplot is shown in Figure 4.4.

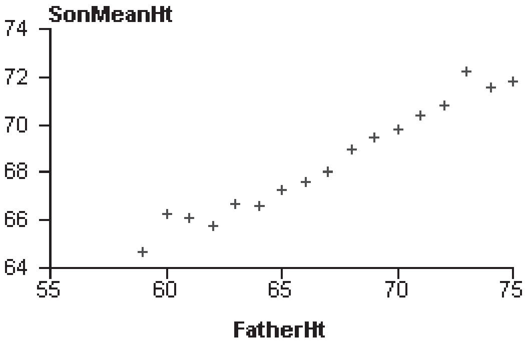


Figure 4.4 A scatterplot of mean son's height against father's height (heights in inches)

From the scatterplot you can see that the mean height of the sons tends to increase with father's height – tall fathers tended to have tall sons, and short fathers tended to have short sons. However, an interesting point emerges from looking more closely at the data matrix and the scatterplot – this is that, although tall fathers did tend to have tall sons, on average the height of the sons was less than the height of the fathers; for example, for fathers 73 inches tall, the average height of their sons was approximately 72.2 inches. Similarly, short fathers had sons who were not, on average, as short as themselves.

It was data similar to those collected by Pearson, but on a smaller scale (only about 200 father–son pairs), that led Sir Francis Galton (1822–1911) in the 1880s to the idea of *regression*. Galton called the phenomenon just described 'regression back to the population mean' or, as he put it, 'toward the mediocre'. You can see from the scatterplot that the sample means lie approximately on a straight line, so it would seem reasonable to model the way the mean height of sons depends on father's height by a linear relationship. In Chapter D4, a method is described for choosing a line to model this relationship. The line obtained is called the *least squares fit line* or the *regression line*.

Obtaining a scatterplot: a summary

A scatterplot is obtained as follows.

◇ Choose **Scatterplot...** from the **Plot** menu.

◇ Select a variable name for the x variable and a variable name for the y variable. The scatterplot is then displayed.

Chapter D2, Section 5
Exploring normal distributions

In this section, you are invited to use *OUStats* to explore the properties of normal distributions. You will need to use **Normal distribution...** from the **Stats** menu for all the activities. In general, you will probably find it quicker and easier to edit values in the boxes than to use the mouse to mark areas under the normal curves.

If necessary, refer to page 102 for an introduction to use of the 'Normal distribution' window.

Activity 5.1 Different means

For each pair of values of the parameters μ and σ in the table below, find the area under the normal curve between the values $a = \mu - \sigma$ and $b = \mu + \sigma$. What do you notice?

μ	σ	a $(= \mu - \sigma)$	b $(= \mu + \sigma)$	Area under curve between a and b
0	1	-1	1	
2.5	1	1.5	3.5	
-3	1	-4	-2	

A solution is given on page 146.

Activity 5.2 Different standard deviations

(a) For each pair of values of the parameters μ and σ in the table below, find the area under the normal curve between the values $a = \mu - \sigma$ and $b = \mu + \sigma$. What do you notice?

μ	σ	a $(= \mu - \sigma)$	b $(= \mu + \sigma)$	Area under curve between a and b
0	1	-1	1	
0	5	-5	5	
0	140	-140	140	

(b) The results obtained in this activity and in Activity 5.1 illustrate a general result for normal distributions. Write down what you think this result might be. Test your conjectured 'result' for a pair of values of μ and σ of your own choice.

Solutions are given on page 146.

Activity 5.3 Two standard deviations from the mean

(a) For each pair of values of the parameters μ and σ in the table below, and for a pair of values of your own choice, find the area under the normal curve between the values $a = \mu - 2\sigma$ and $b = \mu + 2\sigma$. What do you notice?

μ	σ	a $(= \mu - 2\sigma)$	b $(= \mu + 2\sigma)$	Area under curve between a and b
0	1	-2	2	
0	5	-10	10	
20	5	10	30	
20	50			
?	?			

(b) These results illustrate a general result for normal distributions. Write down what you think this result might be. Test your conjectured 'result' for a pair of values of μ and σ of your own choice.

Solutions are given on page 146.

Activity 5.4 Three standard deviations from the mean

For each of the pairs of values of μ and σ in the table in Activity 5.3 (including those of your own choice), find the area under the normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$. Comment on your results.

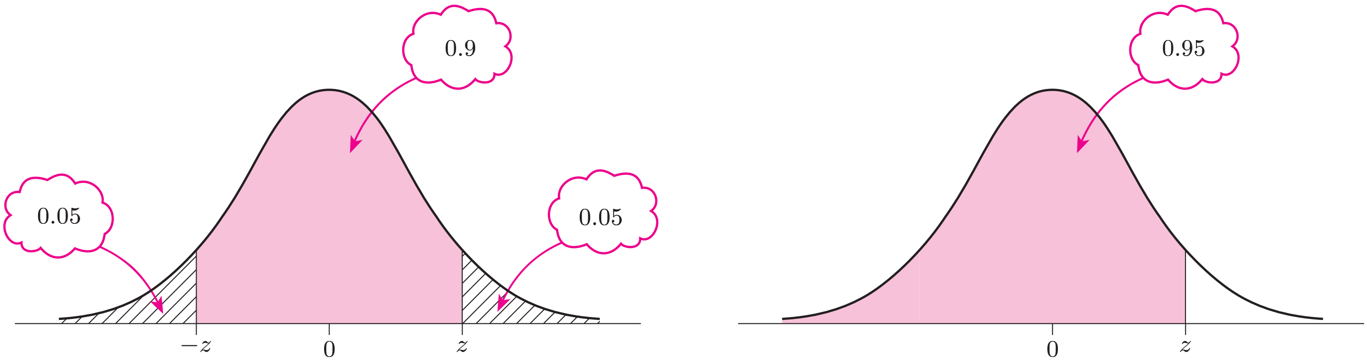
A solution is given on page 146.

Activity 5.5 90% of values

(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Follow the instructions below to find the value z such that the area under the normal curve between $-z$ and z is equal to 0.9.

To do this, you must first turn it into a question that you can answer using the **Normal distribution...** facility. The software allows you to find a value for **A** directly, given the value of **Area to left of A**, so the first thing to do is to calculate the area to the left of z.

The total area under the normal curve is 1, and the curve is symmetrical about the mean, $\mu = 0$; so, if the area between $-z$ and z is 0.9, then the area in each tail (that is, below $-z$ or above z) is equal to $\frac{1}{2} \times 0.1 = 0.05$. This is illustrated in Figure 5.1(a).



(a) The area between $-z$ and z

(b) The area to the left of z

Figure 5.1 Finding the area to the left of z

It follows that the total area to the left of z is 0.95; this is shown in Figure 5.1(b). So put 0.95 in the **Area to left of A** box, ensure that the value in the **Area between A and B** box is no greater than 0.05, and click on **Use areas to calc A & B**; the required value z will appear in the **A** box.

If the area between A and B is given as greater than 0.05 then an error message will appear, since the sum of the two areas cannot exceed 1.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 146.

Activity 5.6 95% of values

(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Find the value z such that the area under the normal curve between $-z$ and z is equal to 0.95.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 146.

Activity 5.7 99% of values

(a) Consider a normal distribution with parameters $\mu = 0$ and $\sigma = 1$. Find the value z such that the area under the normal curve between $-z$ and z is equal to 0.99.

(b) Now consider a normal distribution with parameters $\mu = 20$ and $\sigma = 5$. Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(c) Select values for μ and σ different from those in parts (a) and (b). Find the area under the normal curve between the values $\mu - z\sigma$ and $\mu + z\sigma$, where z is the value that you found in part (a).

(d) Suggest a general result for normal distributions.

Solutions are given on page 147.

Chapter D3, Section 3 *Confidence intervals on the computer*

In Subsection 3.1, you will have the opportunity to use computer simulations to check the interpretation of a 95% confidence interval given in Section 2 of Chapter D3. And in Subsection 3.2, you will learn how to use *OUStats* to calculate a 95% confidence interval for a population mean from a large sample of data.

3.1 Interpreting a confidence interval

In Section 2, it was stated that if a large number of samples is drawn from a population, and a 95% confidence interval for the population mean is calculated from each sample, then approximately 95% of these confidence intervals will contain the population mean μ. In this subsection, you will have the opportunity to investigate the accuracy of this statement. You will be using the *Simulations* package to generate many samples and to calculate the corresponding confidence intervals.

For each different population distribution that you use, you can investigate what proportion of 95% confidence intervals contain the population mean, to ascertain whether it is indeed approximately 95%. The software can be used to simulate taking samples from either a normal distribution or a geometric distribution.

Activity 3.1 Confidence intervals for the mean of a normal distribution

(a) Start the *Simulations* package running as follows.

◇ Click on the **start** menu, move the mouse pointer to **All Programs**, then **MST121**, and finally click on **MST121 Simulations**.

Alternatively, double-click on the **MST121 Simulations** icon on your desktop.

Now click on **Confidence intervals** (either the tab or the panel) to open this simulation.

The option of **Thick lines...** is available from the **Options** menu.

At the top of the window are two buttons, labelled **Normal** and **Geometric**. The default option is **Normal**; this is the option that you will be using in this activity.

(b) The default values of the parameters μ and σ are 0 and 1, respectively; and the default values of the sample size and the number of samples are 25 and 100. Run the simulation with these values to see what happens.

Each confidence interval is represented by a horizontal line segment on the diagram. The population mean μ is marked by a vertical line down the centre of the diagram, and any confidence interval that does not contain μ is displayed in a different colour from those that do contain μ. Thus it is possible to identify easily those intervals that do not contain μ. When the simulation ends, you can scroll back to see how many of the intervals failed to include the population mean. Alternatively, notice that the number of confidence intervals which do contain the population mean μ is displayed in the box at the bottom left of the window.

There are also vertical lines at distances $\sigma/2$, σ and $3\sigma/2$ either side of the mean.

The first time that I ran the simulation, 94 out of the 100 confidence intervals contained the population mean μ; the other six did not. You may well have obtained a different number. Run the simulation several times, and note down the results. On average, approximately what proportion of your confidence intervals contained the population mean μ?

(c) Run the simulation several times for values of the parameters μ and σ of your own choice, and note down your results.

For each pair of values that you used, on average what proportion of the confidence intervals contained the population mean μ?

(d) Now change the sample size to 100 (say). Run the simulation several times for different values of μ and σ of your own choice, and note down your results in each case.

On average, what proportion of the confidence intervals contained the population mean μ?

(e) Investigate the proportion of confidence intervals which contain the population mean μ for further different sample sizes and parameter values. Note down your results.

(f) Comment briefly on your results and on any points you may have noticed about the confidence intervals that you obtained for different sample sizes.

Comment

I ran the simulation ten times for samples of size 25 from a normal distribution with mean 0 and standard deviation 1. I obtained the following results for the number of confidence intervals (out of 100) which contained the population mean μ.

94 92 93 97 94 97 95 97 91 94

That is, 94.4% of all the intervals contained the population mean μ. Another ten simulations produced 93.7% of intervals containing μ. In both cases, the proportion of confidence intervals which contained μ was only a little less than 95%.

For samples of size 100 and $\mu = 50$, $\sigma = 10$, I obtained the following results.

97 94 95 96 94 96 96 97 93 94

Overall, 95.2% of the confidence intervals contained the population mean μ. I used the simulation for various other sample sizes between 40 and 400, and for a number of different parameter values, and in each case obtained similar results: approximately 95% of the confidence intervals contained the population mean μ.

In general, I noticed that (as expected) the confidence intervals were narrower for larger sample sizes. It was observed earlier, in Chapter D2, that the sample standard deviation varies less from sample to sample when the sample size is large than when it is small. So we should expect the width of the confidence intervals to vary less from sample to sample for the larger sample sizes. This was so for my simulations. Did you notice that there was less variation in the width of the confidence intervals for the larger sample sizes that you tried than for the smaller sample sizes?

Activity 3.2 Confidence intervals for the mean of a geometric distribution

Now click on the **Geometric** button. The default value of the parameter p is 0.5.

Use this simulation to investigate the proportion of 95% confidence intervals for the mean of a geometric distribution that actually contain the population mean. Run the simulation for various values of the parameter p and for various sample sizes from 25 upwards.

Recall from Chapter D1 that a geometric distribution with parameter p has mean $\mu = 1/p$. It can also be shown that its standard deviation is $\sigma = \mu\sqrt{1-p}$.

Comment briefly on your results. In particular, write down any points you may have noticed about the lengths of confidence intervals for different sample sizes, and about the proportion of confidence intervals which contain the population mean for different sample sizes.

Comment

I ran the simulation for a range of sample sizes similar to those which I used in Activity 3.1 when investigating confidence intervals for the mean of a normal distribution. I tried several values for the parameter p: $\frac{1}{6}, \frac{1}{2}, \frac{4}{5}, \ldots$. For each parameter value and for each sample size of 50 or larger that I tried, I found that the proportion of confidence intervals that contained the population mean was approximately 95%. However, when I took samples of size 25, the proportion of confidence intervals that contained the population mean was generally a little lower than 95%. For instance, for $n = 25$ and $p = \frac{1}{2}$, just over 91% of my intervals contained the population mean; and for $n = 25$ and $p = \frac{1}{6}$, only about 90% of my intervals contained the population mean. In all cases the proportion was less than 95%.

You may recall that, for large sample sizes, the sampling distribution of the mean may be approximated by a normal distribution, and the approximation improves as the size of the samples increases. A geometric distribution is right-skew (for any value of the parameter p), and the approximation is not nearly as good for samples of size 25 as it is for larger sample sizes. As a result, rather less than 95% of the confidence intervals actually contain the population mean for samples as small as 25.

This effect is seen at its most extreme for values of p close to 1. For example, with samples of size 25 and $p = 0.99$, less than 25% of confidence intervals contain the population mean. (Interpretation of the diagram in this case is an interesting exercise. The value of μ is $1/0.99 \simeq 1.01$, and many confidence intervals consist of the single value 1, corresponding to a sample of 25 values each of which is 1.)

The value of σ is $\sqrt{0.01}/0.99 \simeq 0.1$.

You may also have noticed that the width of the confidence intervals varied greatly for samples of size 25. This occurs because for samples of size 25 from a geometric distribution, the sample standard deviation varies greatly from sample to sample. As for the normal distribution, the sample standard deviation, and hence the width of the confidence intervals, varies much less from sample to sample for larger sample sizes.

3.2 Calculating a confidence interval

In order to calculate a 95% confidence interval for a population mean, the values of the sample mean $\overline{x}$ and the sample standard deviation s need to be calculated. For large samples, this can be a tedious exercise using a calculator, so you were spared carrying out these calculations in Section 2; in each example and activity, you were given the values of $\overline{x}$ and s.

In this subsection, you will not be given these summary statistics. Instead, you will be given the data in a file, and invited to use *OUStats* to calculate confidence intervals. The sample mean and sample standard deviation are calculated automatically when using *OUStats* to find a confidence interval.

In the first activity, you will be shown how to find a confidence interval for the mean height of Cambridge men in 1902. You will be able to check that the calculations agree with those carried out using a calculator in Section 2. You will need to use *OUStats* for all the activities in this subsection.

To start up *OUStats*, click in turn on **start**, **All Programs**, **MST121** and **MST121 OUStats**, or just double-click on the **MST121 OUStats** desktop icon.

Activity 3.3 Finding a confidence interval

The data on the heights of 1000 Cambridge men are contained in the file HEIGHTS.OUS. Open this file now.

A 95% confidence interval for a population mean is obtained using **Confidence interval...** in the **Stats** menu. Click on **Stats**, and choose **Confidence interval...** (by clicking on it). From the list of variables that appears, select 'Height | Frequency', and click on **Select**. The confidence interval is then calculated.

The output includes the sample mean, sample standard deviation and sample size (for information), and a statement of the 95% confidence interval for the population mean. In this case, the confidence interval given is $(68.7128, 69.0312)$. So we can be fairly sure that the mean height of all Cambridge men in 1902 was between 68.71 inches and 69.03 inches.

In Section 2, we obtained $(68.7, 69.1)$ for the 95% confidence interval. The slight discrepancy between these two results is due to rounding error: in Section 2, we used values for the mean and standard deviation which had been rounded to 3 significant figures, whereas *OUStats* calculates the values of the mean and standard deviation to much greater accuracy and uses these values to calculate a confidence interval.

Activity 3.4 Sample size

The procedure described in Section 2 for calculating a 95% confidence interval for a population mean should be used only when the sample size is at least 25. Any results obtained using the formula given there for a smaller sample would be inaccurate and unreliable. In this activity, you are asked to explore what happens if you try to use *OUStats* to calculate a confidence interval for a sample of fewer than 25 items of data.

The data file BAR.OUS contains data on the gross hourly earnings (in pence) in 1995 of a sample of 14 female bar staff. Open the file now, and instruct the computer to calculate a 95% confidence interval for the mean gross hourly earnings in 1995 for female bar staff. What output do you obtain?

Comment

You will have found that, because the sample size is less than 25, the software produces a message telling you that the sample size must be at least 25. However, most statistics packages are not so friendly: if you ask for an inappropriate procedure to be carried out, it will be done. So it is important for you to know when a procedure should or should not be used.

Activity 3.5 Cuckoo eggs

The lengths in millimetres of the 243 cuckoo eggs which were represented in Figure 1.5(d) of Chapter D2 are contained in the file CUCKOOS.OUS. Use these data to find a 95% confidence interval for the mean length of all cuckoo eggs.

Source: O. H. Latter, 'The egg of *Cuculus canorus*', *Biometrika* 1 (1902) pages 164–176.

A solution is given on page 147.

Activity 3.6 Authorship and sentence length

In Activity 2.4, you calculated a 95% confidence interval for the mean sentence length in a book by G. K. Chesterton. The data on sentence lengths are contained in the data file AUTHORS.OUS, together with data on sentence lengths in two other books: *The Work, Wealth and Happiness of Mankind* by H. G. Wells and *An Intelligent Woman's Guide to Socialism* by G. B. Shaw. These data were collected by C. B. Williams in an investigation into sentence length as a criterion of literary style. (Williams chose these particular books for his investigation because, in his words, 'all three deal with sociological subjects and none of them are in the "conversational style"'.)

Source: C. B. Williams, 'A note on the statistical analysis of sentence-length, as a criterion of literary style', *Biometrika* 31 (1940) pages 356–361.

Open the file AUTHORS.OUS now, and explore its contents; remember that you can obtain information about the data in the file using **Notes...** from the **File** menu.

(a) How does the distribution of sentence lengths vary between authors? In order to help you answer this question, obtain frequency diagrams for each of the authors, and compare them.

Hint: Choose **Frequency diagram...** from the **Plot** menu to obtain each of the diagrams in turn. Then choose **Tile** from the **Window** menu, so that you can view all three diagrams together. The scales on the three diagrams will be different.

If you wish, you can adjust the size of a window, and hence the appearance of any scale on a diagram within, by dragging the mouse. Instructions for doing so were given just before Activity 3.3 of Chapter D2 (page 112 in this book).

(b) Compare the mean sentence lengths for the three authors. Does there appear to be a difference? Which author seems to write the longest sentences? Which author seems to write the shortest sentences?

(c) Find a 95% confidence interval for the mean sentence length in each book. What do you conclude from your results?

Solutions are given on page 147.

Activity 3.7 Birthweights

The file BIRTHWT.OUS contains the birthweights of 37 male and 34 female babies, all of whom were born two weeks 'early', that is, at the end of a 38-week gestation period. Find 95% confidence intervals for the mean birthweight of baby boys born two weeks early and for the mean birthweight of baby girls born two weeks early. Comment on your results.

A solution is given on page 148.

Calculating confidence intervals: a summary

A 95% confidence interval for a population mean based on a large (≥ 25) sample from the population is obtained as follows.

◇ Choose **Confidence interval...** from the **Stats** menu.

◇ Select the appropriate variable name(s) from the variable list that appears (to indicate the data to be used), and click on **Select**.

A 95% confidence interval is calculated using each set of data selected.

Chapter D4, Section 2
Exploring the data

In this section, the use of *OUStats* to produce boxplots is illustrated for the data on city block scores which are given in Table 1.1 of Chapter D4. You will be invited to explore the data further to see whether there is a relationship between the time spent memorising the positions of the objects and the score obtained on the test.

Activity 2.1 Obtaining boxplots

The data on city block scores and on memorisation times are contained in the file MEMORY.OUS. Open the file now, and read the information about the data given in **Notes...** .

Boxplots are obtained using the **Plot** menu. Click on **Plot**, and choose **Boxplot...** (by clicking on it). To obtain boxplots for the city block scores of the two groups on the same diagram, select YScore and EScore as follows: click on YScore, then, while holding down the `[Ctrl]` key, click on EScore; both YScore and EScore should now be highlighted. Click on **Select**, and the boxplots will be produced.

You should find that the boxplots look similar to those in Figure 1.4 of Chapter D4, although the five key values are not displayed. When a 'Boxplot' window is open, you can display a list of the five key values by clicking the mouse button while the pointer is within the box or close to either whisker. Try this now.

Next obtain boxplots for the memorisation times of the two groups, YTime and ETime (on a single diagram). Check that they look similar to those in the solution to Activity 1.5(b) of Chapter D4.

In Section 1, we observed that the boxplots for the city block scores suggest that, generally, the young people performed better on the test than the elderly people. However, the boxplots for the memorisation times indicate that the young people spent longer studying the positions of the objects. Does spending longer studying the positions improve performance on the test? If so, then this could explain why the young people performed better on the test.

To investigate whether memorisation time and performance on the test are related, we must look at the city block scores and memorisation times of the individuals who took the test. This information is available in the file MEMORY.OUS. The data in the file are paired. For instance, the first entry in the column headed EScore and the first entry in the column headed ETime relate to one person from the elderly group, and so on.

Activity 2.2 Is performance related to memorisation time?

Obtain two scatterplots, one for the young people and one for the elderly people. Plot memorisation time on the horizontal axis, and city block score on the vertical axis. (Use **Scatterplot...** from the **Plot** menu.) Is there any evidence of a relationship between memorisation time and city block score for either group? Describe any patterns in the scatterplots.

Instructions for obtaining a scatterplot were given in Chapter D2, Activity 4.1 (page 117 in this book).

Comment

You should find that, in both scatterplots, there is a tendency for the city block score to decrease as the memorisation time increases. However, there is a lot of scatter in the plots, so the relationships are weak.

Activity 2.3 Combining the data

It is difficult to tell from the two separate scatterplots whether a young person and an elderly person who spend similar times studying the positions of the objects obtain similar city block scores. We can investigate this by plotting all the data on the same diagram. Obtain a scatterplot for all 27 people who took the test, with memorisation time on the horizontal axis and city block score on the vertical axis. The data for all 27 people are in the columns headed Score and Time.

What can you deduce from the scatterplot? Is there any evidence that young people do better on the test – that is, have lower city block scores – than elderly people who spend a similar length of time memorising the positions of the objects?

Comment

Figure 2.1 contains a scatterplot showing the city block scores and memorisation times for all 27 people who took the test. Different plotting symbols have been used in this figure for the young and the elderly. (It is not possible to use different symbols using *OUStats*.)

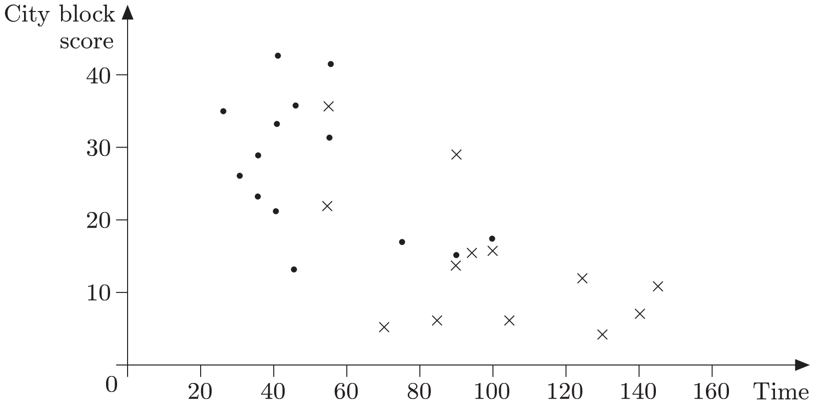


Figure 2.1 A scatterplot of city block scores and memorisation times (× for Young, • for Elderly)

Looking at the scatterplot as a whole, there appears to be a relationship between the time spent studying the positions of the objects and the city block score obtained: in general, the city block score decreases as the memorisation time increases. There is some overlap in the memorisation times of the people in the young and the elderly groups: in each group, there were individuals who spent between 55 and 100 seconds studying the positions of the objects. This part of the scatterplot provides very little evidence that city block scores are lower for young people than for elderly people who spent similar times memorising the positions of the objects.

It seems possible that the better performance of the young group on the test is due to the fact that, in general, they spent longer than the elderly group studying the positions of the objects. Of course, we do not know whether the elderly people would have done as well as the young people if they had all spent the same time studying the positions of the objects.

This will be investigated further in Section 6.

The final activity in this section will give you some practice at obtaining boxplots on the computer and interpreting them. You will obtain further practice in Section 4.

Activity 2.4 Earnings of primary school teachers

The file PRIMARY.OUS contains data on the gross weekly earnings in 1995 of 91 primary school teachers, of whom 54 are women and 37 are men. Obtain boxplots for the earnings of the women and the men. What do the boxplots tell you about the relative earnings in 1995 of male and female primary school teachers?

A solution is given on page 148.

Obtaining boxplots: a summary

Boxplots are obtained as follows.

◇ Choose **Boxplot...** from the **Plot** menu.

◇ Select one or more variable names, and click on **Select**.

◇ To display a list of the five key values on a boxplot, click the mouse button while the pointer is within the box or close to one of the whiskers.

Chapter D4, Section 4
Testing for a difference

In this section, the use of *OUStats* to carry out a two-sample z-test is explained. An essential first step in any investigation is to look at the data. So, in each case, you will be asked to compare the data visually using boxplots before performing a two-sample z-test.

Activity 4.1 Wing lengths of meadow pipits

In this activity, the data from Table 3.1 of Chapter D4 on the wing lengths of male and female meadow pipits will be used to demonstrate the use of *OUStats* to perform a two-sample z-test.

(a) The data are in the file PIPITS.OUS. Open this file now. Compare the wing lengths of the male and female meadow pipits using boxplots. Check that these boxplots agree with those given in Figure 3.1 of the main text in Chapter D4.

(b) The boxplots suggest that there is a difference between the wing lengths of male and female meadow pipits. So now we shall carry out a two-sample z-test to investigate this apparent difference. The first stage is to write down the null and alternative hypotheses: these are

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F,$$

where μ_M, μ_F are the mean wing lengths of the populations of male and female meadow pipits, respectively. (These hypotheses were stated in Section 3.)

The second stage is to calculate the test statistic. This is the part of the hypothesis test that the computer can do for you. Click on **Stats**, and choose **Two sample z-test...** (by clicking on it). You need to specify the data to be used. Select MLength (which contains the wing lengths of the males) as the first variable, and FLength (which contains the wing lengths of the females) as the second variable. Click on **Calc**, and the calculations will be performed.

The output includes the mean, the standard deviation and the sample size of each of the two samples, and the test statistic. According to *OUStats*, the numerical value of z, the test statistic, is 7.5624. Notice that this differs slightly from the value we obtained in Section 3: there we obtained $z = 7.63$. This discrepancy is due to rounding error: in Section 3, to calculate the test statistic, we used values of the means and standard deviations which had been rounded to three significant figures, whereas *OUStats* calculates the means and standard deviations to many more significant figures than this, and then uses these values to calculate the test statistic.

The third and final stage in a hypothesis test is to draw a conclusion (as in Section 3). Since the test statistic z equals 7.5624, which is greater than 1.96, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean wing length of male meadow pipits is not equal to the mean wing length of female meadow pipits. And since the sample mean is greater for the males than for the females, this suggests that the mean wing length of males is greater than the mean wing length of females.

Activity 4.2 Sample sizes

In Section 1, the city block scores on a memory test of a group of 13 young people and 14 elderly people were compared using boxplots. And in Section 2, you reproduced these boxplots using *OUStats*. Since the two-sample z-test depends on the Central Limit Theorem, both sample sizes must be at least 25 for the test to be used. Try using the software to perform the test for the data on city block scores. (The data are in the file MEMORY.OUS.) What happens?

Comment

You will have found that, for these data, the software produces a message telling you the sample sizes and reminding you that both sample sizes must be at least 25. The test is not carried out. This happens whenever either of the sample sizes is less than 25. This is another friendly feature of the software: most statistics packages will carry out your instructions to perform a two-sample z-test whether or not it is an appropriate procedure to use. If it is not appropriate, because the sample sizes are too small, then the results produced using the test would be unreliable and possibly misleading.

Activity 4.3 Authorship and sentence length

In Activity 3.6 of Chapter D3 (page 130 in this book), you explored the distribution of sentence lengths for three books, one by each of three authors – G. K. Chesterton, H. G. Wells and G. B. Shaw. You also obtained confidence intervals for the mean sentence lengths of the three books, and compared them. In this chapter, two methods for comparing samples of data have been described: first boxplots for a visual comparison, and then the two-sample z-test to test for a difference between two population means.

(a) Compare the sentence lengths of the three authors using boxplots (the data are in the file AUTHORS.OUS).

(b) Use the two-sample z-test to investigate whether there is a difference between the mean sentence lengths in the book by G. K. Chesterton and the book by H. G. Wells. State your hypotheses, the test statistic and your conclusion clearly. Note that you should include a statement of your hypotheses, the test statistic and your conclusions in your record of every hypothesis test that you carry out. This applies whether you use a calculator or a computer to carry out the calculations.

(c) Test for a difference between the mean sentence lengths in the book by G. K. Chesterton and the book by G. B. Shaw. Again, state clearly your hypotheses, the test statistic and your conclusion.

Solutions are given on page 149.

Activity 4.4 Birthweights of babies

The file BIRTHWT.OUS contains the birthweights of 37 male and 34 female babies, all of whom were born two weeks 'early', that is, at the end of a 38-week gestation period. In Activity 3.7 of Chapter D3 (page 131 in this book), you found 95% confidence intervals for the mean birthweight of baby boys born two weeks early and for the mean birthweight of baby girls born two weeks early.

(a) Obtain boxplots for the birthweights of the boys and girls. Comment on what they tell you about the birthweights of boys and girls born two weeks early.

(b) Use the two-sample z-test to investigate whether there is a difference between the mean birthweight of boys born two weeks early and the mean birthweight of girls born two weeks early. State clearly your hypotheses, the test statistic and your conclusion.

Solutions are given on page 149.

Activity 4.5 Earnings of primary school teachers

The file PRIMARY.OUS contains data on the gross weekly earnings in 1995 of 91 primary school teachers, of whom 54 are women and 37 are men. In Activity 2.4 (page 134 in this book), you compared the earnings of the men and women using boxplots. Use the two-sample z-test to investigate whether there was a difference between the mean gross weekly earnings (in 1995) of male primary school teachers and female primary school teachers. State clearly your hypotheses, the test statistic and your conclusion.

A solution is given on page 150.

Two-sample z-test: a summary

The two-sample z-test is carried out as follows.

◇ Choose **Two sample z-test...** from the **Stats** menu.

◇ Select two variables, one from each of the two drop-down menus, and click on **Calc**.

The test is carried out only if both sample sizes are at least 25. If either sample size is less than 25, then an error message is produced.

Chapter D4, Section 6
Fitting a line to data

In this section, *OUStats* will be used to calculate the least squares fit line for the concrete data given in Section 5. You will then have the opportunity to investigate the relationships between several other pairs of variables. In each case, you will be asked to obtain a scatterplot and, if it seems appropriate, to fit a straight line to the data and use the equation of this line to make predictions.

Activity 6.1 Finding the least squares fit line

The data on the pulse velocity and crushing strength of concrete, given in Table 5.1 of Chapter D4, are contained in the file CONCRETE.OUS.

(a) Open the file now, and obtain a scatterplot of the data with pulse velocity along the x-axis and crushing strength along the y-axis.

This drop-down menu also provides the opportunity to alter the symbol used for plotting data points.

You can now display the regression line on the scatterplot by opening the drop-down **Options** menu within the 'Scatterplot' window and clicking on 'Regression line on/off'. The resulting display also shows the parameters of the regression line: a slope of 25.8874 and a y-intercept of -87.8283. (The regression line may be removed from the scatterplot by repeating these steps.)

(b) The equation of the least squares fit line can also be obtained using the **Stats** menu. Choose **Regression...** (by clicking on it). Now select 'Velocity' as the first (x) variable and 'Strength' as the second (y) variable, and click on **Calc**. The equation of the least squares fit line is displayed lower in the window, in the form

$$y = -87.8283 + 25.8874x.$$

So the equation of the least squares fit line is $y = -87.83 + 25.89x$, where y is the crushing strength of concrete and x is the pulse velocity for the concrete. This is the equation that was quoted in Section 5.

Activity 6.2 How tall will my son be?

Pearson's data on the heights of 1078 father–son pairs are contained in the file PEARSON.OUS.

(a) Obtain a scatterplot of son's height against father's height (with father's height along the x-axis), and then add the least squares fit line to the plot.

(b) Obtain the equation of the least squares fit line, and use it to predict the height of the son of a 70-inch-tall man.

(c) By referring to the scatterplot you obtained in part (a), comment on how precise you think this estimate might be.

Solutions are given on page 150.

When will Old Faithful erupt?

Every year, tourists flock to the Yellowstone National Park in Wyoming in the United States. One of the attractions is the Old Faithful geyser, which erupts about 20 times a day, on average. As you saw in Chapter D2, the eruptions vary in length, the shortest lasting just over a minute and the longest about 5 minutes. The intervals between eruptions also vary a lot. Sometimes the waiting time from the end of one eruption to the beginning of the next is as short as 40 minutes, but it can be as long as an hour and a half. Unlucky visitors can have a long wait! So is there a way of predicting when the next eruption will occur, so that visitors can be informed?

See Chapter D2, Activity 1.2.

In August 1978, the geyser was observed between 6 am and midnight on eight consecutive days; the duration of each eruption and the waiting time until the next eruption were both recorded. The purpose of collecting the data was to investigate whether the duration of one eruption could be used to predict when the next is likely to occur. The question was: ‘Is there a relationship between the duration of an eruption and the waiting time until the next eruption?’ And if there is, can we use the data to formulate a rule for predicting when the next eruption is likely to occur?

Activity 6.3 Exploring the relationship

The data on the eruptions of the Old Faithful geyser in August 1978 are contained in the file FAITHFUL.OUS.

(a) Obtain a scatterplot with the duration of an eruption along the x-axis and the waiting time until the next eruption along the y-axis.

(b) What does the scatterplot tell you about the relationship between the duration of an eruption and the waiting time until the next eruption? Do you think a straight line would be a suitable model for the relationship?

(c) Obtain the equation of the least squares fit line, and use it to predict the waiting time until the next eruption following eruptions which last for the following times.

(i) 1.5 minutes (ii) 3 minutes (iii) 4.5 minutes

(d) Add the least squares fit line to your scatterplot (if not already done). Comment on how accurate you think your predictions are.

Solutions are given on page 151.

In fact, when the data were collected, an error was made in recording one day’s results: the eruption times and intervals between eruptions were paired incorrectly. This meant that there were quite a number of anomalous points which did not fit the general pattern that you observed. So it was thought that a useful prediction rule could not be formulated. The error was discovered only several years later.

Memory and age

In Section 1 of Chapter D4, an investigation into spatial memory in the young and elderly was discussed. Two groups of people, one young and one elderly, tackled a memory test in which eighteen everyday objects were placed on a 10 by 10 square grid. After a person had studied the positions of the objects for as long as they wished, the objects were removed. Then they were asked to replace the objects in the same positions. Two pieces of data were noted for each person: the time spent studying the positions of the objects, and a measure of accuracy of recall – the city block score.

The city block score is described in Chapter D4, Activity 1.1.

Activity 6.4 Does performance improve with time?

In Section 2 of Chapter D4, you used *OUStats* to investigate whether performance on the memory test was related to the time spent memorising the positions of the objects. The data are in the file MEMORY.OUS.

You obtained the three scatterplots required in Activity 6.4 previously, in Activities 2.2 and 2.3. However, the possibility of fitting a line to the data did not form part of those activities.

(a) For the elderly group, obtain a scatterplot of city block score against memorisation time. Is there any evidence of a relationship between the time spent memorising the positions of the objects and performance on the test? If you think it appropriate, fit a line to the data (with memorisation time as the explanatory variable).

(b) Now obtain a similar scatterplot for the young group. Comment on the relationship between the time spent memorising the positions of the objects and performance on the test. If you think it appropriate, fit a line to the data.

(c) Now obtain a scatterplot for the data for the two groups combined, and fit a line to the data. According to this model, what is the predicted score of a person whose time spent memorising the positions of the objects is as follows?

Note that the memorisation times are recorded in seconds.

(i) 1 minute (ii) 2 minutes (iii) 3 minutes

Comment briefly on your results.

Solutions are given on page 151.

Activity 6.5 Comparing the fit lines

Obtain a printout of the scatterplot for the two groups combined (without the least squares fit line on it). On this scatterplot, draw the two fit lines whose equations you found in parts (a) and (b) of Activity 6.4 – one for the elderly group and one for the young group. Comment briefly on what you deduce from this diagram.

A solution is given on page 152.

Finding the equation of the regression line: a summary

The equation of the least squares fit line, or the regression line of y on x, can be obtained as follows.

- ◇ Choose **Regression...** from the **Stats** menu.
- ◇ Select a variable name for the first (x) variable and a variable name for the second (y) variable, and click on **Calc**.

Alternatively, the parameters of the equation are given along with the least squares fit line that can be added to a scatterplot, as below.

Scatterplots: a summary

A scatterplot is obtained as follows.

- ◇ Choose **Scatterplot...** from the **Plot** menu.
- ◇ Select a variable name for the x variable and a variable name for the y variable. The scatterplot is then displayed.

The least squares fit line may be included on a scatterplot as follows.

- ◇ Open the **Options** drop-down menu within the 'Scatterplot' window.
- ◇ Choose 'Regression line on/off' from the list of options (by clicking on it). The regression line appears, along with the values of the parameters for its equation.

The regression line may be removed by repeating these steps.

The plotting symbol on a scatterplot may be changed as follows.

- ◇ Open the **Options** drop-down menu within the 'Scatterplot' window.
- ◇ Choose the symbol that you require from the list that appears (by clicking on the corresponding 'Plot points as' row).

Appendix: Entering and editing data

Creating a new data file

- Open *OUStats*. Choose **New...** from the **File** menu. You will then see a dialogue box for entering the numbers of rows and columns that you want to have available. The default size of grid is 200 rows by 40 columns. If you need more rows or columns than this, edit the appropriate box. Then click on **OK** or press `[Enter]`.
- You will then see a data window showing a data matrix of the size that you specified in the dialogue box. To enter a column of data in, say, column V1, click on the first row of V1 and type in the first value. Press `[Enter]` to move the cursor to the next row, and type in the next value. Press `[Enter]` again, and continue entering values in this way. To type in a second column of data, click on the first row of that column and repeat the entering of values as for the first column. Note that if you prefer to enter data across the rows rather than down the columns, you can simply press `[Tab]` instead of `[Enter]` after typing each value. This will move the cursor to the right.
- You can go back and edit any value by clicking on it, using `[Backspace]` to delete, typing in the new value, and pressing `[Enter]`. You can also move around the data matrix using the arrow keys.

Naming and renaming columns

When a new data window is selected, the default names for the columns (that is, variables) are V1, V2, V3, etc. To rename a column (once some data have been entered), choose **Rename column...** from the **Edit** menu, select the original name in the dialogue box which appears (by clicking on the corresponding editing box), type in the new name, and click on **OK**.

Saving a new data file

Choose **Save As...** from the **File** menu. You should then see a 'Save' dialogue box which shows a scrollable list of data (.OUS) file names.

Enter a name for your data file. Then click on **Save**. Your file is now saved in the same folder as all the other data files, and should be listed with them when you choose **Open...** from the **File** menu.

Entering frequency data

Frequency data must be entered with the values in one column (in V1, say) and the corresponding frequencies in another column (in V2, say). To designate the values in a column as values for frequency data, place the mouse pointer in the heading for that column (V1 in this case) and click with the *right* mouse button. Then choose (for this example) 'Set frequency: V2' from the menu that appears. The values in column V2 are then designated as frequencies for the values in column V1. (You can check this by clicking again with the right mouse button in the heading for column V1. You will notice that the top (highlighted) line of the resulting menu now reads 'Frequency: V2' in place of 'No frequencies'.)

You must enter all values in the columns and, if needed, rename the columns, before designating values as frequency data.

When the values in one column have been designated in this way as frequencies for another column, the names of the two columns are linked together in dialogue boxes which display variable names. For instance, 'Height | Frequency' refers to the frequency data which have been entered into the two columns labelled 'Height' and 'Frequency', then linked together as described above.

Making your own Notes file

When you create a new data file using *OUStats*, it is a good idea to record information about the file (the source of the data, an explanation of the variable names, etc.) in an associated Notes file. You can do this as follows.

◇ Open the data file (if it is not already open).

◇ Click on **Notes...** in the **File** menu.

◇ Type your notes about the data file into the 'Notes' window.

◇ To save the notes, click on **Save** in the **File** menu (whether or not the 'Notes' window itself is open).

The commands **Cut**, **Copy** and **Paste**, in the **Edit** menu, are available whenever a 'Notes' window is open.

Transferring data into and out of OUStats

Instructions for transferring data from another application into *OUStats*, or vice versa, are given in the *OUStats* Help file, which may be accessed via the *OUStats* **Help** menu.

Solutions to Activities

Chapter D2

Solution 3.1

Frequency diagrams for the four data sets are shown in Figure S2.1.

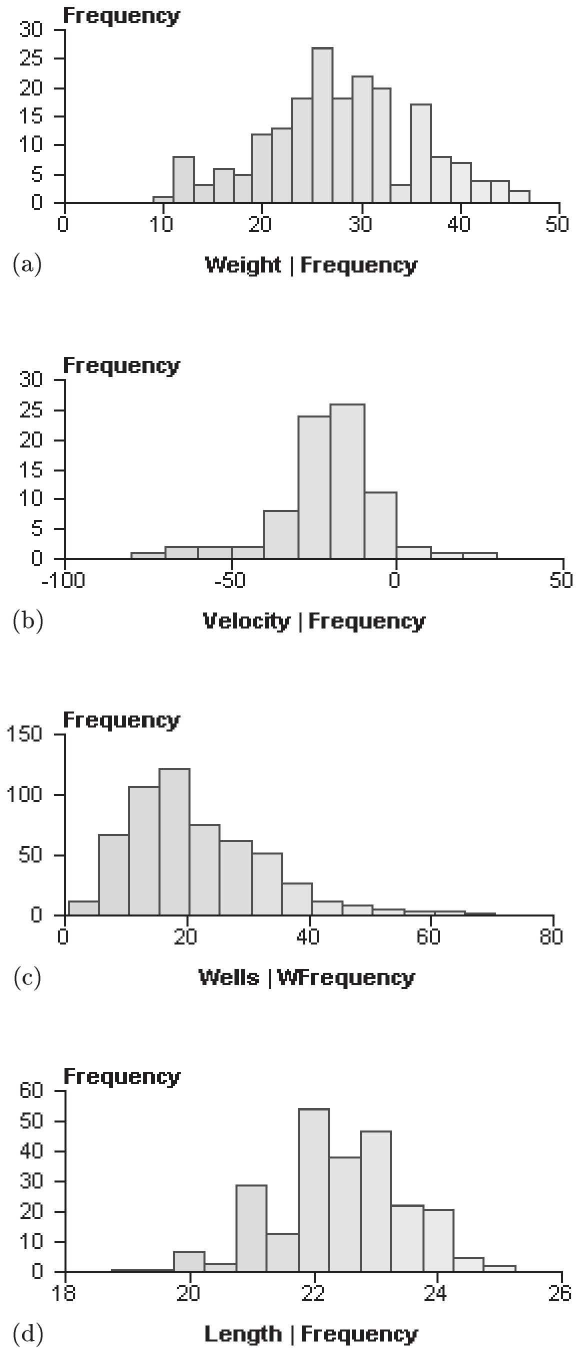


Figure S2.1 Four frequency diagrams

(a) The first frequency diagram (for the weights of Irish dipper nestlings) was obtained using the first interval starting value and interval width suggested in the activity (9 and 2, respectively).

(b) For the frequency diagram for the radial velocities, a first interval starting value of -80 and an interval width of 10 were used, corresponding to the grouping of the data.

(c) To obtain the frequency diagram for the lengths of sentences written by H. G. Wells, a first interval starting value of 0.5 and a width of 5 were used. This means that the first bar represents sentences of lengths 1 to 5 inclusive, the second bar represents sentences of lengths 6 to 10 inclusive, and so on. You may well have chosen different values for the start of the first interval and the interval width.

(d) The lengths of cuckoo eggs are given to the nearest half millimetre, so a length recorded as 19 mm could be anywhere between 18.75 mm and 19.25 mm. To obtain the frequency diagram shown, a first interval starting value of 18.75 and an interval width of 0.5 were used.

The frequency diagram for the lengths of sentences written by H. G. Wells is right-skew (the right tail is longer than the left tail), so a normal distribution is not an appropriate model in this case. The other three frequency diagrams are all roughly symmetrical with a single clear peak, so a normal model is worth considering for these data. You are asked to investigate these three data sets further in the next three activities.

Solution 3.3

I fitted a normal model with mean -21 and standard deviation 16. (As in Activity 3.2, I used one significant figure more for the parameters of the normal distribution than are given in the data.)

The four frequency diagrams obtained are shown in Figure S2.2. (The size of each random sample was 80, the same as the size of the sample of data.)

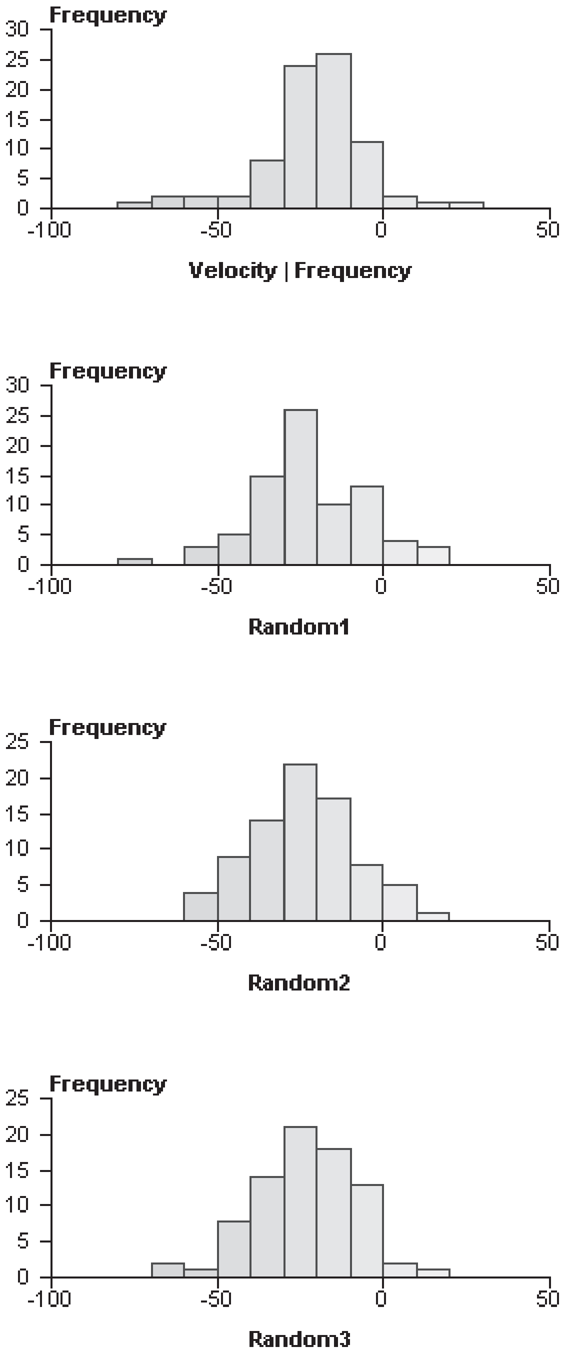


Figure S2.2 Four frequency diagrams

There is considerable variation between the shapes of the frequency diagrams for the random samples. However, they all appear to be less sharply peaked than the frequency diagram for the data, which has longer shallower tails. It would appear that a normal model may not be a very good fit for the data. Certainly, the fit does not seem to be as good in this case as the fit of the normal curve to the weights of Irish dipper nestlings in Activity 3.2.

Solution 3.4

I fitted a normal distribution with mean 22.4 and standard deviation 1.08. (Again, I used one significant figure more for the parameters of the normal distribution than are given in the data.)

The four frequency diagrams obtained are shown in Figure S2.3. (The size of each random sample was 243, the same as the size of the sample of data.)

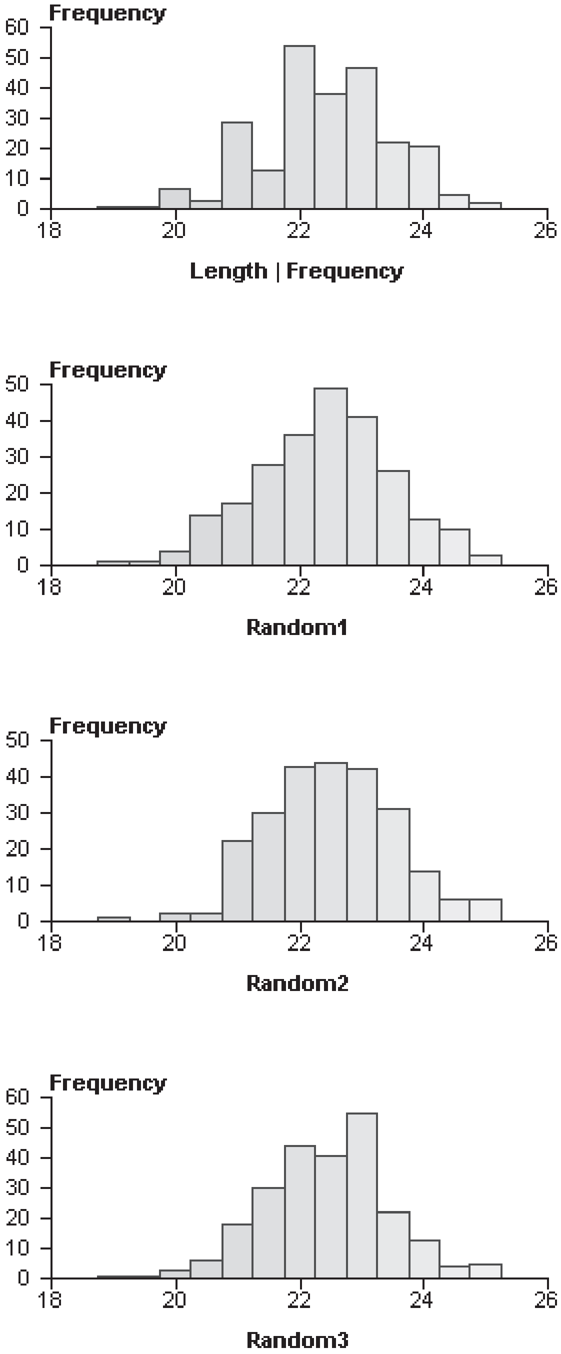


Figure S2.3 Four frequency diagrams

Again, there is considerable variation between the shapes of the frequency diagrams for the random samples. The main difference between the frequency diagram for the data and the frequency diagrams for the random samples is that the frequency diagram for the data is more jagged. Apart from this, the fit seems to be quite good.

Solution 5.1

All the areas between a and b are equal to 0.683 (to 3 s.f.).

Solution 5.2

(a) All the areas between a and b are equal to 0.683 (to 3 s.f.).

(b) The area under a normal curve from one standard deviation below the mean to one standard deviation above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within one standard deviation of the mean is about 68.3% for any normal distribution.

Solution 5.3

(a) All the areas between a and b are equal to 0.954 (to 3 s.f.).

If you have *OUStats* set to display 6 significant figures (the default), then it will show these areas as 0.9545 (suppressing two trailing zeros). However, to 8 s.f. the areas are displayed as 0.954 499 74, giving 0.954 to 3 s.f.

(b) The area under a normal curve from two standard deviations below the mean to two standard deviations above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within two standard deviations of the mean is about 95.4% for any normal distribution.

Solution 5.4

In this case, all the areas between a and b are equal to 0.997 (to 3 s.f.).

The area under a normal curve from three standard deviations below the mean to three standard deviations above the mean is the same whatever the mean and standard deviation of the distribution. In fact, the proportion of values within three standard deviations of the mean is about 99.7% for any normal distribution.

Solution 5.5

(a) The area to the left of 1.644 85 is 0.95, so the area between −1.644 85 and 1.644 85 is equal to 0.9. The required value is $z = 1.644\,85$ (to 6 s.f.).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 1.644\,85 \times 5 = 11.775\,75$ and $\mu + z\sigma = 20 + 1.644\,85 \times 5 = 28.224\,25$ is equal to 0.9.

(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 1.644\,85\sigma$ and $\mu + 1.644\,85\sigma$ is equal to 0.9.

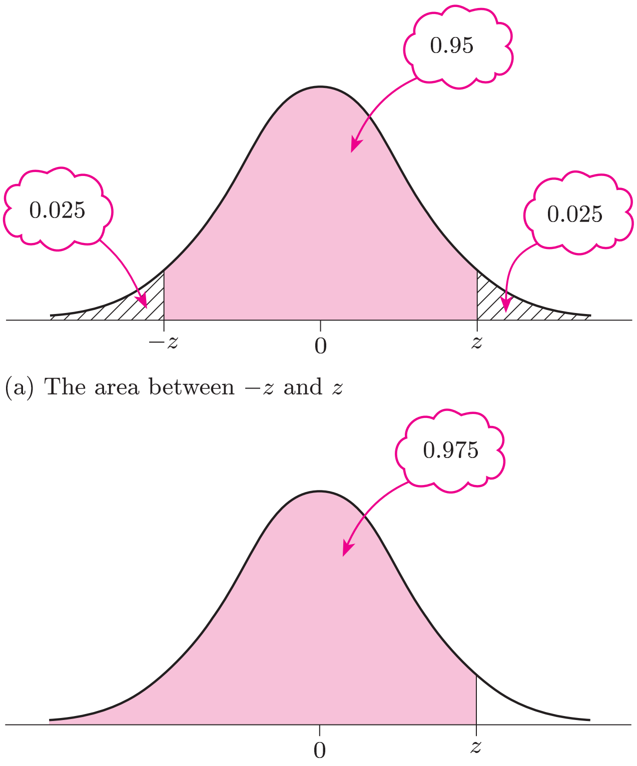
(d) The area under a normal curve within 1.644 85 standard deviations of the mean is 0.9, whatever the values of the mean and standard deviation.

In practice, we shall normally use the value of z calculated to only 3 significant figures (which is 2 decimal places in this case). Rounding to 2 decimal places gives 1.64, and this is the value that is commonly used; we do not usually require greater accuracy than this. So we have the following result.

For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 90% of the population are within 1.64 standard deviations of the mean, that is, between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$.

Solution 5.6

(a) If the area between $-z$ and z is equal to 0.95, then the area of each tail is $\frac{1}{2} \times 0.05 = 0.025$. So the total area to the left of z is 0.975. This is illustrated in Figure S2.4.



(a) The area between $-z$ and z

(b) The area to the left of z

Figure S2.4 Finding the area to the left of z

The area to the left of 1.959 96 is equal to 0.975. The required value is $z = 1.959\,96$ (to 6 s.f.).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 1.959\,96 \times 5 = 10.2002$ and $\mu + z\sigma = 20 + 1.959\,96 \times 5 = 29.7998$ is equal to 0.95.

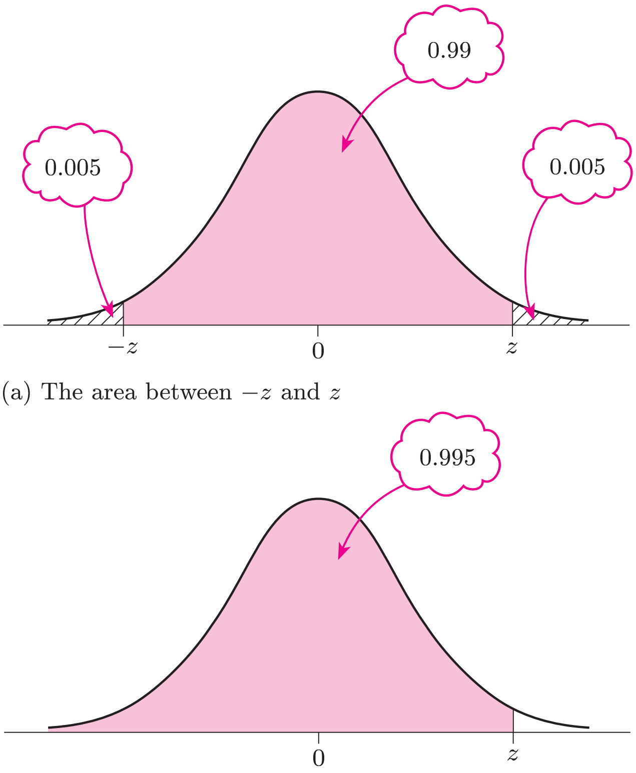
(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 1.959\,96\sigma$ and $\mu + 1.959\,96\sigma$ is equal to 0.95.

(d) The area under a normal curve within $1.959\,96 \simeq 1.96$ standard deviations of the mean is 0.95, whatever the values of the mean and standard deviation.

For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 95% of the population are within 1.96 standard deviations of the mean, that is, between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Solution 5.7

(a) If the area between $-z$ and z is equal to 0.99, then the area of each tail is $\frac{1}{2} \times 0.01 = 0.005$. So the total area to the left of z is 0.995. This is illustrated in Figure S2.5.



(a) The area between $-z$ and z

(b) The area to the left of z

Figure S2.5 Finding the area to the left of z

For the area to the left of z to be 0.995, z must be equal to $2.575\,83$ (to 6 s.f.).

(b) The area under the normal curve between $\mu - z\sigma = 20 - 2.575\,83 \times 5 = 7.120\,85$ and $\mu + z\sigma = 20 + 2.575\,83 \times 5 = 32.879\,15$ is equal to 0.99.

(c) Whatever values of μ and σ you choose, you should find that the area under the normal curve between $\mu - 2.575\,83\sigma$ and $\mu + 2.575\,83\sigma$ is equal to 0.99.

(d) The area under a normal curve within $2.575\,83 \simeq 2.58$ standard deviations of the mean is 0.99, whatever the values of the mean and standard deviation.

For a population modelled by a normal distribution with mean μ and standard deviation σ, approximately 99% of the population are within 2.58 standard deviations of the mean, that is, between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$.

Chapter D3

Solution 3.5

The 95% confidence interval for the mean length (in millimetres) of cuckoo eggs given by *OUStats* is $(22.2762, 22.5469)$. So, rounding to 3 significant figures, a 95% confidence interval for the mean length (in millimetres) of cuckoo eggs is $(22.3, 22.5)$.

Solution 3.6

(a) Frequency diagrams of sentence lengths for each of the three authors are shown in Figure S3.1. In each case, a first interval starting value of 0.5 and an interval width of 5 were used.

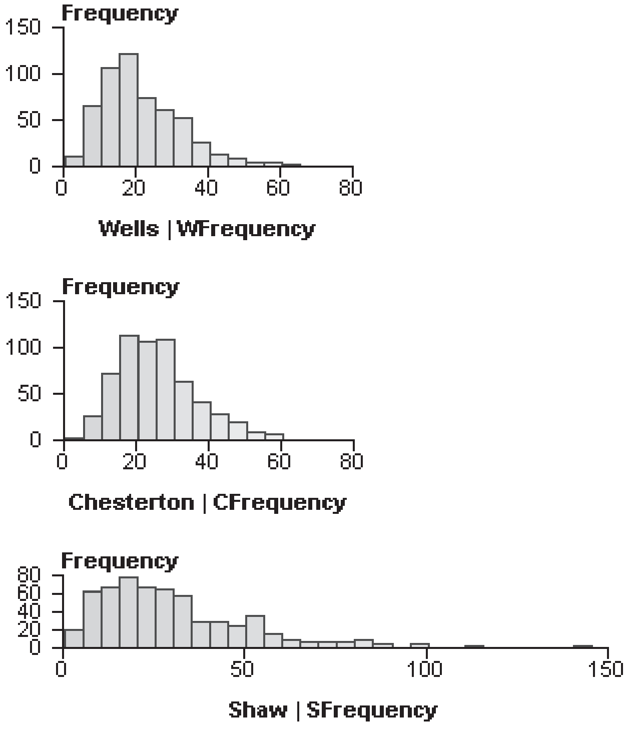


Figure S3.1 Frequency diagrams of sentence lengths

In order to produce these diagrams, which have different scales, I used **Tile** from the **Window** menu and then adjusted the sizes of the windows by dragging the mouse.

The sentence lengths from the book by Shaw are much more variable than the sentence lengths from either of the other two books. There are more long sentences, and some very long ones.

(b) The mean sentence lengths for Wells, Chesterton and Shaw are (according to *OUStats*) 21.6799, 25.6131 and 31.1642, respectively. After rounding to 3 s.f., these are 21.7, 25.6 and 31.2. So, on average, Shaw seems to have written the longest sentences and Wells the shortest.

(c) The 95% confidence intervals for the mean sentence lengths in each of the three books, as given by *OUStats*, are as follows.

Wells	$(20.7435, 22.6164)$
Chesterton	$(24.7502, 26.4759)$
Shaw	$(29.4516, 32.8767)$

Rounding the confidence limits to 3 significant figures (one more than is given in the data) gives the following.

Wells	$(20.7, 22.6)$
Chesterton	$(24.8, 26.5)$
Shaw	$(29.5, 32.9)$

These confidence intervals are represented in the sketch in Figure S3.2.

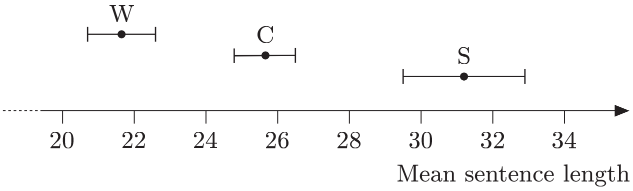


Figure S3.2 Confidence intervals

Since these confidence intervals do not overlap, this suggests that the mean sentence lengths in the books are different. It looks as though the mean sentence length in the book by Shaw is greater than the mean sentence length in the book by Chesterton, and that this is in turn greater than the mean sentence length in the book by Wells.

However, we cannot say how confident we are that the means are different. For instance, although we are 95% confident that the mean length of sentences in the book by Wells is between 20.7 and 22.6, and we are 95% confident that the mean length of sentences in the book by Chesterton is between 24.8 and 26.5, we cannot put a figure to our confidence that the two means are different: this might be more or less than 95%, but we cannot say what it is simply by comparing the two 95% confidence intervals.

If we want to be able to quantify our confidence that the means are different, then a different approach is needed; a method which compares the two samples of data is required, rather than a method which looks at each sample separately. Such a method is discussed in the next chapter.

Solution 3.7

The 95% confidence intervals for the mean birthweights (in grams) of boys and girls born two weeks early, as given by *OUStats*, are as follows.

Boys	$(3068.36, 3349.53)$
Girls	$(2912.11, 3239.48)$

These confidence intervals are represented in the sketch in Figure S3.3.

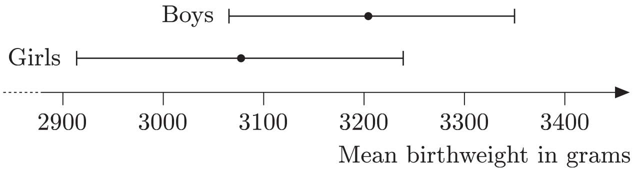


Figure S3.3 Confidence intervals

Although the confidence limits are higher for the boys than for the girls, the two intervals overlap, so we cannot draw conclusions from these confidence intervals about whether there is a difference between the mean birthweights of boys and girls born two weeks early.

Chapter D4

Solution 2.4

Boxplots of the men's and women's earnings are shown in Figure S4.1.

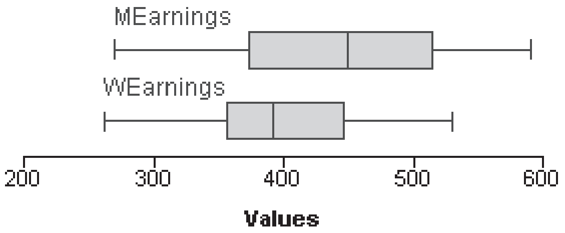


Figure S4.1 Gross weekly earnings (in pounds) of male and female primary school teachers

The earnings of the men seem to be generally higher than the earnings of the women, although the difference does not appear to be great. The difference is greatest for the highest earners in the two groups, and is very small for the lowest earners in the two groups.

Solution 4.3

(a) The boxplots are shown in Figure S4.2.

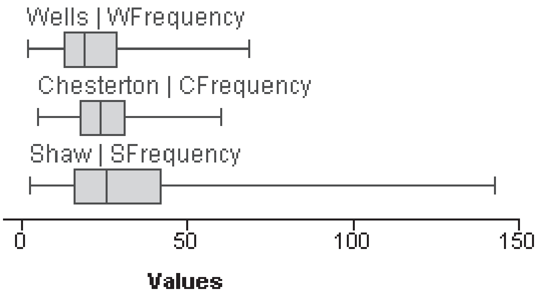


Figure S4.2 Boxplots of sentence lengths

As you can see, all the books contained quite a high proportion of fairly short sentences; and all three boxplots are right-skew, indicating that long sentences are less common than shorter ones in all three books. The main difference between the sentence lengths in the three samples seems to be that some of the sentences in the book by Shaw are longer than any of the sentences in the samples from the other two books. The average sentence length, as measured by the median, is longest for the sample from the book by Shaw and shortest for the sample from the book by Wells.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_C = \mu_W,$$
$$H_1 : \mu_C \neq \mu_W,$$

where μ_C is the mean sentence length in the book by G. K. Chesterton, and μ_W is the mean sentence length in the book by H. G. Wells.

The test statistic (obtained from *OUStats*) is $z = 6.054\,18$.

Since the test statistic $z = 6.054\,18 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean sentence length in the book by Chesterton is not equal to the mean sentence length in the book by Wells. The sample mean is greater for the book by Chesterton than for the book by Wells, so this suggests that the mean sentence length is greater for the book by Chesterton than for the book by Wells.

(c) The null and alternative hypotheses may be written as

$$H_0 : \mu_C = \mu_S,$$
$$H_1 : \mu_C \neq \mu_S,$$

where μ_C is the mean sentence length in the book by G. K. Chesterton, and μ_S is the mean sentence length in the book by G. B. Shaw.

The test statistic (obtained from *OUStats*) is $z = -5.673\,78$.

Since the test statistic $z = -5.673\,78 < -1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that the mean sentence length in the book by Chesterton is not equal to the mean sentence length in the book by Shaw. The sample mean is greater for the book by Shaw than for the book by Chesterton, so this suggests that the mean sentence length is greater for the book by Shaw than for the book by Chesterton.

Solution 4.4

(a) The boxplots are shown in Figure S4.3.

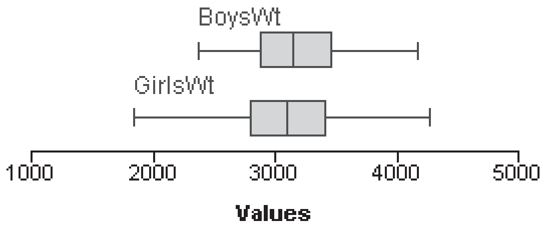


Figure S4.3 Boxplots of birthweights

No great difference is apparent between the birthweights of the boys and the girls, although the median birthweight is slightly higher for the boys than for the girls, and there is less spread in the boys' birthweights.

(b) The null and alternative hypotheses may be written as

$$H_0 : \mu_B = \mu_G,$$
$$H_1 : \mu_B \neq \mu_G,$$

where μ_B is the mean birthweight (in grams) of baby boys born two weeks early, and μ_G is the mean birthweight of baby girls born two weeks early.

The test statistic (obtained from *OUStats*) is $z = 1.209\,51$.

Since $-1.96 < z < 1.96$, we cannot reject the null hypothesis at the 5% significance level. The data do not provide sufficient evidence at the 5% significance level to reject the hypothesis that the mean birthweight of boys born two weeks early is equal to the mean birthweight of girls born two weeks early.

Solution 4.5

The null and alternative hypotheses may be written as

$$H_0 : \mu_M = \mu_F,$$
$$H_1 : \mu_M \neq \mu_F,$$

where μ_M is the mean gross weekly earnings (in pounds) of male primary school teachers in 1995, and μ_F is the mean gross weekly earnings of female primary school teachers in 1995.

The test statistic (obtained from *OUStats*) is $z = 2.790\,34$.

Since the test statistic $z = 2.790\,34 > 1.96$, we reject the null hypothesis at the 5% significance level in favour of the alternative hypothesis. We conclude that there was a difference between the mean gross weekly earnings in 1995 of male and female primary school teachers. The sample mean is greater for the men than for the women, so this suggests that the mean gross weekly earnings of male primary school teachers in 1995 was greater than the mean gross weekly earnings of female primary school teachers in 1995.

Solution 6.2

(a) The scatterplot and the least squares fit line are shown in Figure S4.4. Note that each 'plus' on the scatterplot may represent the heights of one father–son pair or of many pairs: it is not possible to tell how many, as frequencies are not represented. If you feel that the line shown does not look as if it is the best fit line, then this is probably because when looking at the scatterplot you cannot take into account the relative frequencies of the different pairs of values.

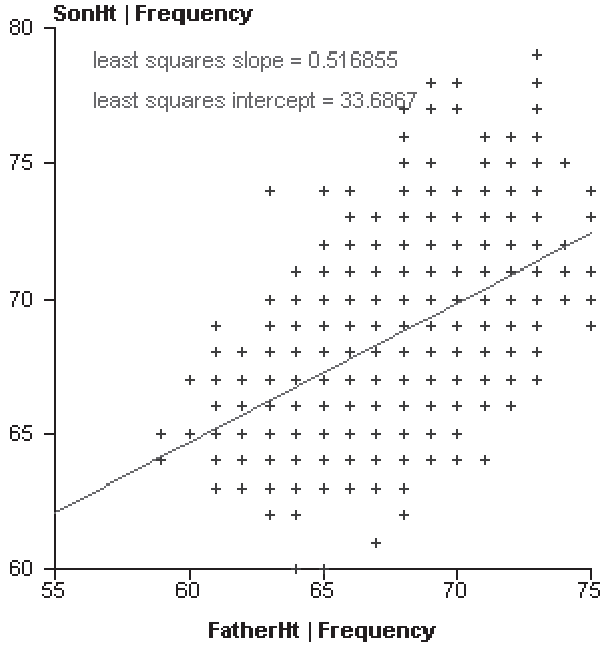


Figure S4.4 A scatterplot of son's height against father's height

(b) The equation of the least squares fit line is

$$y = 33.69 + 0.5169x,$$

where y represents son's height and x represents father's height (giving the parameters to 4 s.f.).

The predicted height of the son of a 70-inch-tall man is

$$y = 33.69 + 0.5169 \times 70 \simeq 69.9 \text{ inches.}$$

However, the data on which the model is based were collected in the 1890s for fathers and sons in the UK. If the average height of men has continued to increase from generation to generation, possibly by different amounts in different generations, then data collected now might well lead to a slightly different model. This prediction applies only to the sons of fathers in the UK in the 1890s. Moreover, the families measured in Pearson's study were predominantly middle class, so the prediction applies to middle class families in the 1890s. (A different model might have been required for the heights of fathers and sons in working class families.)

(c) There is a lot of scatter in the plot, so any individual son of a 70-inch-tall man could be a lot taller or shorter than 69.9 inches. This height is an estimate of the mean height of sons of 70-inch-tall men in the UK in the 1890s.

Solution 6.3

(a) The scatterplot is shown in Figure S4.5.

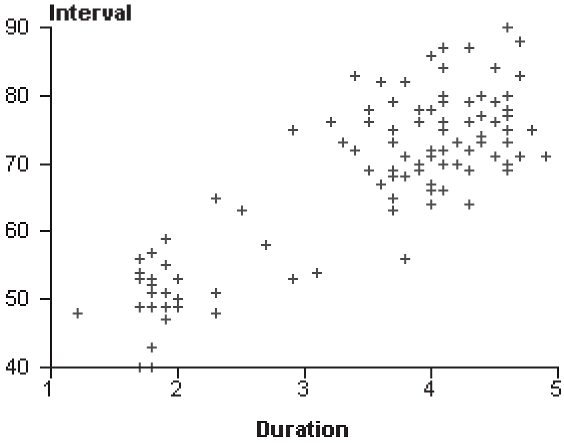


Figure S4.5 A scatterplot of time to next eruption against duration of eruption

(b) There seem to be two main groups of points on the scatterplot, corresponding to 'short' and 'long' eruptions. 'Short' eruptions lasted two minutes or less; 'long' eruptions lasted more than three minutes. Very few of the eruptions were of intermediate duration. However, overall, the time to the next eruption appears to increase with the length of the current eruption. Following a long eruption, there is a longer wait, on average, until the the next eruption than following a short eruption. However, there is a lot of scatter in the plot, so the relationship is not a very strong one. It does look as though a straight line might summarise the relationship quite well.

(c) The equation of the least squares fit line is

$$y = 33.65 + 9.806x,$$

where x minutes is the duration of an eruption and y minutes is the waiting time until the start of the next eruption.

(i) After an eruption of length 1.5 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 1.5 \simeq 48.4 \text{ minutes.}$$

(ii) After an eruption of length 3 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 3 \simeq 63.1 \text{ minutes.}$$

(iii) After an eruption of length 4.5 minutes, the predicted time until the start of the next eruption is

$$33.65 + 9.806 \times 4.5 \simeq 77.8 \text{ minutes.}$$

(d) The predictions are estimates of the *mean* waiting time until the next eruption following eruptions of lengths 1.5, 3 and 4.5 minutes, and there is a lot of scatter about the regression line. So the next eruption may be much sooner or much later than predicted. Nevertheless, the predictions could be used to give a very rough indication of when the next eruption is likely to occur: the mean waiting time is nearly half an hour longer following a 4.5-minute eruption than following a 1.5-minute eruption. This is a situation where a confidence interval might be more useful than a simple prediction. The lower confidence limit might be useful as an indication of the earliest time that the next eruption is likely to occur.

Solution 6.4

(a) There is a lot of scatter in the plot for the elderly group, but it does look as though there might be a weak relationship between memorisation time and city block score. The equation of the least squares fit line is

$$y = 36.81 - 0.1793x,$$

where x is the memorisation time in seconds, and y is the city block score.

(b) Again there is a lot of scatter in the plot for the young group, but less than in the scatterplot for the elderly group. The equation of the least squares fit line is

$$y = 32.61 - 0.1836x.$$

(c) The equation of the least squares fit line for the combined group is

$$y = 38.07 - 0.2264x.$$

According to this model, the predicted city block scores are as follows.

(i) When the memorisation time is 1 minute, the predicted score is

$$38.07 - 0.2264 \times 60 \simeq 24.$$

(ii) When the memorisation time is 2 minutes, the predicted score is

$$38.07 - 0.2264 \times 120 \simeq 11.$$

(iii) When the memorisation time is 3 minutes, the predicted score is

$$38.07 - 0.2264 \times 180 \simeq -3.$$

Clearly, a city block score of -3 is impossible; the lowest possible score is 0, for someone who replaces all the objects in the correct positions. Note that three minutes is outside the range of memorisation times for the people taking the test, so using the model for prediction is not valid in this case. Clearly, the model is not appropriate for times as long as three minutes. Since the lowest possible score is 0, a model which does not predict values below zero might be considered. Might a curve rather than a straight line provide a more useful model?

Solution 6.5

In order to draw the lines, I found the coordinates of two points on each line: $(0, 36.8)$ and $(150, 9.9)$ for the fit line for the elderly group; $(0, 32.6)$ and $(150, 5.1)$ for the fit line for the young group. The two least squares fit lines are drawn on the scatterplot in Figure S4.6.

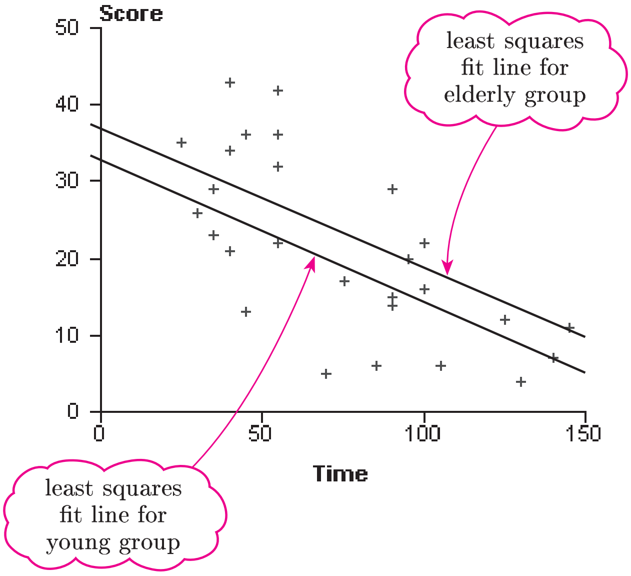


Figure S4.6 The two least squares fit lines

As you can see, the gradients of the two lines are roughly equal. However, the line for the elderly group is a little higher than the line for the young group, suggesting that elderly people do not perform quite as well as young people who spend similar times memorising the positions of the objects. Is there a real difference, or is the observed difference between the lines simply due to sampling variation? This is the sort of question that more advanced regression techniques can tackle. It is possible to fit parallel lines to two data sets and then to carry out a hypothesis test of whether there is a 'real' difference between the intercepts or whether the observed difference might be due to chance. However, we shall not be discussing how to do so in MST121.

Acknowledgements

Grateful acknowledgement is made to the following sources for permission to reproduce material in this book and on the MST121/MS221 Software CD.

Draper, N. R. and Smith, H. (1966), *Applied Regression Analysis*, reprinted by permission of John Wiley and Sons, Inc. All rights reserved.

Fowler, J. and Cohen, L. (1996), *Statistics for Ornithologists, BTO Guide 22*, Second Edition, British Trust for Ornithology.

Reprinted with permission from Johnson, M. P. and Raven, P. H. (1973), 'Species number and endemism: the Galápagos Archipelago revisited', *Science*, **179**, pp. 893–895, Copyright © 1973 American Association for the Advancement of Science.

Lindgren, B. W. and Berry, D. A. (1981), *Elementary Statistics*, Macmillan Publishing Co., Inc., by permission of Prentice-Hall, Inc.

Mazess, R. B., Peppler, W. W. and Gibbons, M. (1984), 'Total body composition by dual-photon (^{153}Gd) absorptiometry', *American Journal of Clinical Nutrition*, **40**, pp. 834–839.

Metzger, W. H. (1935), *Journal of the American Society of Agronomy*, **27**, p. 653, American Society of Agronomy.

Selvin, S. (1991), *Statistical Analysis of Epidemiological Data*, New York, Oxford University Press.

Snedecor, G. W. and Cochran, W. G. (1980), *Statistical Methods*, Seventh Edition, The Iowa State University Press.

Sokal, R. R. and Rohlf, F. J. (1981), *Biometry*, Second Edition, W. H. Freeman and Company Publishers.

Spiegel, M. R. (1972), *Schaum's Outline of Theory and Problems of Statistics in SI Units*, reproduced with permission of The McGraw-Hill Companies.

Terman, L. M. (1919), *The Intelligence of School Children*, Copyright © 1919 by Houghton Mifflin Company. Renewed 1946. Reprinted with the permission of the author's estate.

Trumple, R. J. and Weaver, H. F. (1953), *Statistical Astronomy*, p. 194, Table 1.5, Copyright © 1953 The Regents of the University of California.

Wetherill, G. B. (1981), *Intermediate Statistical Methods*, Chapman and Hall.

Index for OUStats